九州工業大学学術機関リポジトリ
Kyutacar
Kyushu Institute of Technology Academic Repository

Development of Complementary Elementary ModeAnalysis for Integration of Heterogeneous Biological Data into a Complex Metabolic Network

著者	Badsha Md. Bahadur
year	2015-03
その他のタイトル	多様な生物学的データを複雑な代謝ネットワークに統合するための補完的エレメンタリーモード解析法の開発
学位授与年度	平成26年度
学位授与番号	17104甲情工第299号
URL	http://hdl.handle.net/10228/5458

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Development of Complementary Elementary Mode Analysis for Integration of Heterogeneous Biological Data into a Complex Metabolic Network.

Md. Bahadur Badsha

Graduate School of Computer Science and System Engineering, Department of Bioscience and Bioinformatics, Kyushu Institute of Technology 680-4 Kawazu, Iizuka, Fukuoka, 820-8502, Japan.

March 2015

DEDICATED

TO

Almighty Allah

Preface

This dissertation is submitted for the partial fulfillment of the degree of doctor of philosophy. It is based on work carried out between 2012 and 2014 in Systems and Synthetic Biology, Metabolic Engineering and Bioinformatics research group at Professor Kurata's Labolartory, Graduate School of Computer Science and System Engineering, Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, under the supervision of Professor Dr. Hiroyuki Kurata.

Md. Bahadur Badsha
Kyushu Institute of Technology, Japan.

List of Publications

This thesis is based on the following publications:

I. **Md. Bahadur Badsha,** Ryo Tsuboi and Hiroyuki Kurata: Complementary elementary modes for fast and efficient analysis of metabolic networks, *Biochem. Eng. J.*, **90**, 121–130, 2014.

II. **Md. Bahadur Badsha** and Hiroyuki Kurata: Integration of omics into metabolic flux distribution by complementary elementary mode analysis for large-scale metabolic networks, *169th OMICS Group Conference, 3rd International Conference and Exhibition on Metabolomics & System Biology*, March 24-26, 2014, San Antonio, USA.

III. **Md. Bahadur Badsha**, Ryo Tsuboi and Hiroyuki Kurata: Complementary elementary mode analysis for large-scale metabolic networks. *IPSJ SIG Technical Report, BIO-*35 (5), 1-2, 2013, Hokkaido, Japan.

IV. **Md. Bahadur Badsha**, Nusrat Jahan, Md. Nurul Haque Mollah and Hiroyuki Kurata: Metabolic Engineering for Systematic Organization and Analysis of Complex Metabolic Networks, *International Conference on Statistical Data Mining for Bioinformatics, Health, Agriculture and Environment,* 21-24 December, 2012, Bangladesh.

Additional publications not included in this thesis:

V. **Md. Bahadur Badsha,** Nusrat Jahan, Md. Nurul Haque Mollah and Hiroyuki Kurata: Robust Complementary Hierarchical Clustering for Gene Expression Data Analysis by β-Divergence, *J. Biosci. Bioeng*., **116**(3), 397-407, 2013.

VI. **Md. Bahadur Badsha,** Nusrat Jahan, Md. Nurul Haque Mollah and Hiroyuki Kurata: Sequential Extraction of Several Gene-sets with Proper Groups of Individuals by Gene Expression Data Analysis, *International Conference on Statistical Data Mining for Bioinformatics, Health, Agriculture and Environment***,** 21-24 December, 2012, Bangladesh.

Table of Contents

Abstract

Systems biotechnology is an approach to develop comprehensive and ultimately predictive models of how components of a biological system reproduce its observed behavior. The major human diseases like as diabetes, obesity, high blood pressure, cardiovascular disease and cancer are involved in failure of human metabolic systems. Therefore, metabolism is an important biological process, but these are complex and highly interconnected each others. Metabolic network maps are represented by a complex chain of chemical reactions and are highly associated between genes, proteins and enzymes; consequently mathematical and/or computational approaches are necessary for integration of them. Heterogeneous biological data, including genome, transcriptome, proteome, and metabolome are integrated into a pathway-based metabolic model to predict a flux distribution of genetically modified cells under particular conditions. The integration of heterogeneous biological data and model building have become essential activities in biological research as technological advancements continue to empower the measurement of biological data of increasing diversity and scale. But the challenge becomes how to integrate this data to maximize the amount of useful biological information that can be extracted.

Metabolic pathway analysis is theoretically effective in integrating heterogeneous biological data into metabolic network and to offer great opportunities for studying functional and structural properties of metabolic pathways. Metabolic pathway analysis has focused on two approaches, namely, elementary modes (EMs) and extreme pathways (Expas). EM analysis is potentially effective in integrating transcriptome or proteome data into metabolic network analyses and a minimal set of reactions that can maintain the steady state level, while Expa analysis is a subset of EM that contains two additional conditions and one of them condition to make all Expas systematically independent. The EM coefficients (EMCs)

indicate the quantitative contribution of their associated EMs and that can be estimated by maximizing as a particular objective function.

A serious problem of EM/ Expa analysis is that the computational time increases exponentially with an increase in network sizes, which makes the computation of the all EMs/Expas expensive and impracticable for large- or genome-scale networks. Another major problem is that many organisms still does not have provide any specific objective biological function for estimating the EMCs to predict the flux distribution relate to the optimum physiological states and EMs can be described by different scalar products or many possible vectors of each EM, but the predicted flux distributions must be independent of them.

To address such aforementioned problems, in this thesis we present a fast and efficient algorithm, called complementary EM (cEM) analysis, to reduce the number of EMs/Expas. To achieve the computational time improvement, we employ the EM decomposition method that explores major EMs or linear combinations of them which are responsible for the metabolic flux distributions. Flux balance analysis (FBA) is used to generate many possible ranges of metabolic flux distributions as the input data, which is necessary for the EM decomposition method. The maximum entropy principle (MEP) is used as an objective function for estimating the coefficients of cEMs, to renounce the scalar product problem of EMs. MEP is widely used for flux prediction in particular cases where no biological objective function is available and most advantages that it does not depend on the scalar product of each EM.

To demonstrate the feasibility of cEM analysis, we compared it with EM/Expa analysis by using a simulation study with an artificial metabolic network model and real metabolic network analysis by two medium-scale metabolic network model of *E. coli* and a genome scale model for head and neck cancer cells. The cEM analysis greatly reduces the

number of EM, computational time and memory cost for the genome-scale metabolic network. Application of cEM analysis to Genetic Modification of Flux (GMF) accurately predicts the flux distributions of genetic mutants under particular conditions. Use of cEMs analysis, to plans a genetic engineering strategy for genome-scale metabolic network model for producing useful compounds.

Keywords: Systems biotechnology; Integrating biological data; Constraint-based metabolic modeling; Large-scale metabolic network; Elementary mode decomposition; Complementary elementary mode analysis; Quantitative contributions; Prediction speed and accuracy.

Chapter 1

Introduction

1.1 Introduction

In the 21st century, next-generation sequencers enabled the human genome to be decoded at a surprisingly high speed. Furthermore, the production of biological data has become more rapid, owing to the development of systemic high-throughput technology. The crucial question has raised how you understand a huge, complicated biological network and ongoing challenge is to bridge the gaps in our understanding of processes at the molecular, cellular, and tissue levels. It is almost impossible to intuitively understand the behavior of the cellular metabolic networks, due to the complexity of the metabolic pathway interactions and large number of components are involved in the networks.

Systems biotechnology is an approach to develop comprehensive and ultimately predictive models of how components of a biological system reproduce its observed behavior of the metabolic network. Mathematical modeling has been established successfully when applied to relatively small-scale systems, while applications to the large-scale models are being challenged by the practical advances that generate high-dimensional and high-throughput data (Waters et al., 2012). The cellular metabolic and regulatory networks are often large and complex, the construction and analysis of their computational models are can be useful for identifying physiological states and evaluating the effects of network perturbations on desired phenotypes. Recently, genome-scale computational models have gained increasing prominence and importance; capturing stoichiometric model with

thermodynamic constraints have been published for over 30 organisms ranging from relatively simple prokaryotes such as *Escherichia coli* (*E. coli*) (Milne et al., 2009), to complex eukaryotes such as *Homo sapiens* (Duarte et al., 2007 and Ma et al., 2007; Shlomi et al., 2007).

Metabolomics is a branch of system biology, which has been applied to identify and quantize all metabolites in an organism sample under specified living conditions. Strictly speaking metabolism refers all metabolites in an organism or cell. Metabolism is essential in the processes of life. It is mainly consisted of catabolism and anabolism, anabolism refers the process that organisms transfer absorbed nutrients from the external environment into their own components and stores the energy; catabolism on the other hand, refers the process that organisms decompose itself, produce energy then excrete the end products from the decomposition. These processes with the necessary enzymes produce all of the major constituents of the cell.

Metabolic network is an abstract expression of cell metabolism that maps all biochemical reactions into a network for a cell or organism, each metabolite is a node and the reactions are the links or pathways between metabolites which connect the nodes to form a network. This network reflects the interactions between all compounds as well as the enzymes, which involved in the metabolic processes. Metabolic network analysis is a successful way of predicting the metabolic phenotype of an organism under its metabolic genotype and particular conditions which could provide us a better understanding of cellular metabolic processes and the evolution of life. Therefore, predicting the functions of a metabolic network become one of the most important tasks now days (Ma and Zenf, 2003; Patil and Nielsen, 2005 and Wang, 2011).

1.2 Reviews of the Literatures

The biochemical reactions which illustrate various portions of the metabolism are depicted using a metabolic network. In metabolism, constraint-based metabolic modelling methods are systematized biochemical, genetic and genomic knowledge into a mathematical framework that enables a mechanistic description of metabolic physiology. Over the 30 years, the use of constraint-based approaches have evolved, and an increasing significant number of studies have recently combined models with high-throughput data sets for prospective experimentation. These studies have demonstrated the endorsement of increasingly important and relevant biological predictions (Bordbar et al., 2014). With the growing interest of better understandings of the biochemical network, the experimental techniques have improved significantly, however, it is still not powerful enough to determine the whole network or too expensive to conduct, hence some alternative estimation methods have been proposed by metabolic pathway analysis.

Constraint-based models have become a fundamental tool to study genome-scale metabolic networks (Edwards and Palsson, 2000). Such models use governing constraints to restrict potential cellular behavior. The range of all possible behaviors, which is mathematically described by the steady-state flux. Constraint-based analyses are used for predicting the intracellular metabolic fluxes from the metabolic network map in steady-state levels by integrating of the experimental data from genomics, transcriptomics, proteomics, metabolomics, and fluxomics, which are determined by high-throughput technologies. Two possible approaches for constraint-based analysis, namely (1) optimization-based and (2) pathway-based, when the metabolic network is available, as will be illustrated in **chapter 2**. Several optimizations-based approaches have been developed that allow computing, in the altered network, behaviors optimizing a particular network function (Papin et al., 2004). Mathematically, this requires are defining a hypothetical objective function.

In principle, optimization-based methods can be used to analyze the metabolic network and the cellular phenotype of either (1) wild-type or (2) mutant cell. The optimization-based analysis methods, e.g., Flux Balance Analysis (FBA) (Fell and Small, 1986; Varma and Palsson, 1994b), Regulatory-FBA (rFBA) (Covert et al., 2001), Minimization Of Metabolic Adjustment (MOMA) (Segre et al., 2002), Regulatory On/Off Minimization (ROOM) (Shlomi, et al., 2005), Flux Variability Analysis (FVA) (Mahadevan and Schilling, 2003). Optimization-based metabolic analysis with mutant (knock-out, knock-in) cell is obtained using genetic modification techniques, e.g., OptKnock (Burgard et al., 2003), RobustKnock (Tepper and Shlomi, 2009), OptReg (Pharkya and Maranas, 2006) OptGene (Patil et al., 2005), OptORF (Kim and Reed, 2010) and OptForce (Ranganathan et al., 2010).

The metabolic pathway maps are complex in every living cell, where a coherent set of enzymes that catalyzes by various biochemical reactions (Croes et al., 2005; King et al., 2005 and Zaho et al., 2013). Pathway-based methods are capable of characterizing the entire solution space of the possible metabolic network states without imposing the cellular objective biological function bias. Metabolic flux analysis (MFA) is used for the quantitative estimation of intracellular metabolic fluxes through metabolic pathways and the elucidation of cellular physiology in steady-state level (Stephanopoulos, et al., 1998). Pathway-based analyses are generally employed a constraint-based modeling approach (Price, et al., 2004), e.g., FBA that uses a stoichiometric matrix and an objective function to define a network's allowable solution space. The target metabolic flux capacity is provided by optimizing a specific objective function such as cell growth, energy, biomass, adenosine triphosphate (ATP) production or metabolite synthesis (Papin et al., 2004; Raman and Chandra, 2009).

Metabolic pathway analysis has focused on two approaches, namely, elementary modes (EMs) (Schuster et al., 1999) analysis and extreme pathways (Expas) (Schilling et al., 2000) analysis. The EM analysis is potentially effective in integrating transcriptome or proteome data into metabolic network analyses and a minimal set of reactions that can maintain the steady state level, while the Expa analysis is a subset of EM that contains two additional conditions and one of them condition to make all Expas systematically independent. The EM analysis allows one to systematically enumerate all independent minimal pathways that are stoichiometrically and thermodynamically feasible and to offer great opportunities for studying functional and structural properties of metabolic pathways (Stelling et al., 2002; Schwender et al., 2004; Carlson and Srienc, 2004). The EM-based algorithms, e.g., control effective flux (CEF), modified control effective flux (mCEF), enzyme control flux (ECF) and genetic modification of flux (GMF), are very effective in correlating transcriptome or proteome data to their associated metabolic network building or flux distributions and some of those augment metabolic network with the gene regulatory network or enzyme activity profile (Stelling et al., 2002; Cakir et al., 2007; Kurata et al., 2007; Zhao and Kurata, 2009 a b, 2010).

To find the whole set of EMs, distributed memory parallelization and parallel processing have been merged together with compression of the stoichiometric matrix (Jevremovic et al., 2011; Jevremovic and Boley, 2012) or with the remove of biological infeasible solutions (Jungreuthmayer et al., 2013). On the other hand, alternative approaches have been presented without enumerate the whole set of EMs (de Figueiredo et al., 2009; Ip et al., 2011; Machado et al., 2012). The EM decomposition method (Ip et al., 2011) has been developed to pick up the major EMs or linear combinations of EMs, which are responsible for the metabolic flux distributions for metabolic networks, while the entire flux distributions must be input for EM

decomposition. The EM coefficients (EMCs), which indicate the quantitative contribution of their associated EMs and which can be estimated by maximizing a particular objective function. The EMs can be described by many possible vector or scalar products of each EM, but the predicted fluxes must be consistent with respect to all of them.

The linear programming (LP) method is often used to predict the metabolic flux distribution, where the maximum biomass and specific metabolite formation are selected as an objective function (Gayen and Venkatesh, 2006). Such objective functions relate to the optimum physiological states, but they are not provided for many organisms. Maximum biomass or ATP production are can vary between different organisms and physiological conditions. Thus that objective functions are not the best choice. The quadratic programming (QP) could optimize EMCs by defining the objective function as the minimal norm of the EMCs, but QP has neither a physical nor a biological background and is still restricted to relatively small-scale networks (Schwartz and Kanehisa, 2005). A serious problem of QP that it depends on the scalar products of each EM. Therefore, the QP method may not be valid for optimizing the EMCs (Badsha et al., 2013). The enzyme control flux linear programing denoted as ECFLP (Kurata et al., 2007), which maximizes and minimizes each EMC to represent its available ranges in the same manner of the α-spectrum method (Wiback et al., 2004), averaging all the estimated EMCs. The ECFLP is practically valuable, but it has neither a biological nor a theoretical background.

To obtain reliable EMCs, the maximum entropy principle (MEP) algorithm (Kurata et al., 2007; Zhao and Kurata, 2009ab, 2010) have been proposed, which is a universal principle established based on Shannon entropy (Shannon, 1948) when insufficient information is available. MEP is widely used for flux prediction in particular cases where no biological objective function is available and most advantages that it does not depend on the scalar

product of each EM. The MEP enthusiastically optimizes hundreds of thousands of the EMCs in large-scale networks.

1.3 Statement of the Problems

The major human diseases like as diabetes, obesity, high blood pressure, cardiovascular disease and cancer are involved in failure of human metabolic systems. Therefore, metabolism is an important biological process, but these are complex and highly interconnected each others. So, it is very important to properly understand the metabolic pathway map in network level. Metabolic network maps are represented by a complex chain of chemical reactions and are highly associated between genes, proteins and enzymes; consequently mathematical and/or computational approaches are necessary for integration of them. Heterogeneous biological data, including genome, transcriptome, proteome, and metabolome are integrated into a pathway-based metabolic model to predict a flux distribution of genetically modified cells under particular conditions. The integration of heterogeneous biological data and model building have become essential activities in biological research as technological advancements continue to empower the measurement of biological data of increasing diversity and scale. But the challenge becomes how to integrate this data to maximize the amount of useful biological information that can be extracted.

A serious problem arises of EM/Expa analysis is that the computational time increases exponentially with an increase in network sizes, which makes the computation of the all EMs/Expas expensive and infeasible for large-scale or genome-scale networks (Segre et al., 2002; Haus et al., 2008; Acuna et al., 2009; Trinh et al., 2009). For example, central metabolism of *E. coli* model, with 112 reactions has more than two million EMs. When possible substrates are extended, the number of EMs increases to more than 26 million (Terzer and Stelling, 2008). Thus, the huge computation time and memory storage are

required to enumerate all the EMs/Expas of large-scale or genome-scale metabolic networks. Another problem is that many organisms still do not provide any specific objective biological function for estimating the EM coefficients (EMCs) to predict the metabolic flux distribution relate to the optimum physiological states. The EMs can be described by different scalar products of each EM, but the predicted fluxes must be independent of them.

1.4 Objectives of the Study

To overcome the aforementioned problems of the existing EM/Expa methods, we develop a fast and efficient algorithm, named complementary EM (cEM) analysis, to reduce the EMs/Expas (Badsha et al., 2014). To enhance the computing time improvement, we employ the EM decomposition method that explores the major EMs or linear combinations of them, which are responsible for the metabolic flux distributions. FBA is used to generate many possible ranges of metabolic flux distributions as the input data necessary for the EM decomposition method. The maximum entropy principle (MEP) is employed as an objective function for estimating the coefficients of cEMs to avoid the scalar product problem of EMs. To demonstrate the feasibility of cEM analysis, we compared it with EM/Expa analysis by using a simulation study with an artificial metabolic network model, and real metabolic network analysis by two medium-scale metabolic network model of *E. coli* and a genome scale model for head and neck cancer cells. The cEM analysis greatly reduces the computational time and memory cost, without generating a full set of EMs nor any biological objective function, which exposing a new window for large-scale metabolic network analysis.

1.5 Layout of the Study

This section contains an overview / structure of the PhD thesis as given below:

The present chapter provides an introduction, reviews of the literatures, a statement of the problems of the study of the existing or ordinary methods, objective of the study to overcome those problems and the layout of the PhD thesis.

In chapter 2, we describe the details about background regarding the heterogeneous biological data integration into metabolic networks. To enhance the quality of the thesis, we discuss details of the metabolic network analysis for the constraint-based method of optimization-based and pathway-based metabolic network analysis, which are used for predicting the steady-state intracellular fluxes from the metabolic network by integrating of the experimental data from genomics, transcriptomics, proteomics, metabolomics, and fluxomics, those are determined by high-throughput technologies.

In chapter 3, we discuss details about the materials and methods of the study. To overcome the problems of the existing methods, we introduce a new concept to develop a fast and efficient algorithm, complementary EM (cEM) analysis, which is critically useful for the integration of heterogeneous biological data into a complex metabolic network map.

In chapter 4, we will show the results compared with the existing methods by synthetic and real metabolic network analysis. To investigate the performance and applicability of the cEM method in a comparison of the existing method, we consider a simulation study by an artificial metabolic network model and real metabolic network analysis by two medium-scale metabolic network model of *E. coli* and a genome scale model for head and neck cancer cells.

In chapter 5, we give the conclusion and scope of this PhD study, and future research interest.

Chapter 2

Background

2.1 Introduction

A central challenge in the development of systems biology is the integration of heterogeneous biological data into metabolic network model. Mathematical and/or computational approaches are needed to not only integrate this heterogeneous biological data, but also use this data to heighten the predictive capabilities of computational models (Blazier and Papin, 2012). With the advent of high-throughput technologies, heterogeneous biological data types have provided quantitative data for thousands of cellular components across a variety of scales. Cellular metabolism is defined as the essential physical and chemical processes for the maintenance of life. A metabolic network for a specific organism contains all metabolic reactions occurring within the living cells of an organism. With the rapid development of genomics and various successful genome projects, biologist deciphered the genome sequence of many organisms and the metabolic networks for such organisms can be faithfully reconstructed from available genome information. Thus the analysis of metabolic network has become essential in further studies, such analysis could help us to achieve a better understanding of the topology and biological functions of different organisms, hence enable us to utilize cellular metabolic process to assist the development of fermentation technology, medical industry and agriculture (Wang, 2011).

2.2 Metabolic Engineering

Metabolic engineering provides clear and insightful information regarding the activity of metabolic complex reaction networks from an individual reaction based perspective. Metabolic engineering is potentially used for systematic organization and analysis of complex metabolic networks (Badsha et al., 2012). The final goal of metabolic engineering is to be able to produce valuable substances on an industrial scale in a cost effective manner. In the earlier times, to increase the productivity of a desired metabolite, a microorganism was genetically modified by chemically induced mutation, and the mutant strain that over-expressed the desired metabolite was then chosen. However, one of the central problems with this technique was that the metabolic pathway affecting the metabolite production was not analyzed, and as a result, constraints to production and relevant pathway enzymes to be modified were unknown (Voit and Torres 2002).

In 1990s, a new technique called metabolic engineering emerged to overcome the problem of the traditional techniques. Metabolic engineering is emerging the directed improvement of product formation or cellular properties through the modification of specific biochemical reactions or the introduction of new ones with the use of recombinant deoxyribonucleic acid (DNA) technology (Stephanopoulos et al., 1998). The successful applications of metabolic engineering are the following some examples, such as, (i) Identification of constraints to lysine production in Corynebacterium *glutamicum* and the insertion of some new genes to relieve these constraints to improve production (Stephanopoulos et al., 1998). (ii) The engineering of a new fatty acid biosynthesis pathway, called reversed beta oxidation pathway which can potentially be catalytically converted to chemicals and fuels (Dellomonaco, 2011). (iii) Improved production of 3-deoxy-D-arabino-

heptulosonate 7-phosphate (DAHP), an aromatic metabolite produced by *E. coil* that is an intermediate in the production of aromatic amino acids (Patnaik and Liao, 1994).

2.3 Systems Biology

Systems biology is the study that tries to understand as a system consisting of a biological molecule of a large number of organisms. It is an interdisciplinary area which uses mathematical and computational models to describe the cellular behavior and phenomena, as well as the overall properties of the biological systems in general (Klipp et al., 2009). In the study of systems biology, the biological complex is considered as a whole, as divergent to studying individual components and interactions. However, this can be a very difficult task for the researchers, even for simple bacteria, and to address the difficulty various modeling techniques have been proposed approaching the enormous complexity at different levels. A wealth of experimentally obtained genomic, transcriptomic, proteomic and metabolomic data can be used to model and simulate either the functioning of the entire biological cell or just one of its segments (Jevremovic, 2013). One of the attracting aims of systems biology is to build-up a mathematical modeling and discover emergent properties, such as, properties of cells, tissues and organisms functioning as a system whose theoretical description is only possible using mathematical techniques which fall under the concern of systems biology (Ahmed, 2008). In the systems biology study, a mathematical or computational approaches are necessary to integrate heterogeneous biological data, such as transcriptome, proteome, metabolome, and fluxome, to build comprehensive metabolic models. Mathematical models are consistently improved or modified by accommodating new experimental data with the current models, enhancing the validation of metabolic networks or the prediction of their dynamic behaviors (Wiback, et al., 2004; Borodina and Nielsen, 2005). Generating a new knowledge through modelling and integration of experimental data in order to develop a holistic understanding of organisms that are studied in systems biology areas. Figure 2.1

illustrates the workflow commonly referred to as the systems biology cycle based on Kitano (2002a,b). With the growing interest of better understandings of the biochemical networks, the experimental techniques were improved significantly, especially by the application to biological systems, however, it is still not powerful enough to determine the whole network or too expensive to conduct, hence some alternative estimation methods have been proposed. Metabolic network analysis successfully predicts the metabolic phenotype which gives us a good idea of what is happening in an organism and how the organisms work under different external environmental conditions.

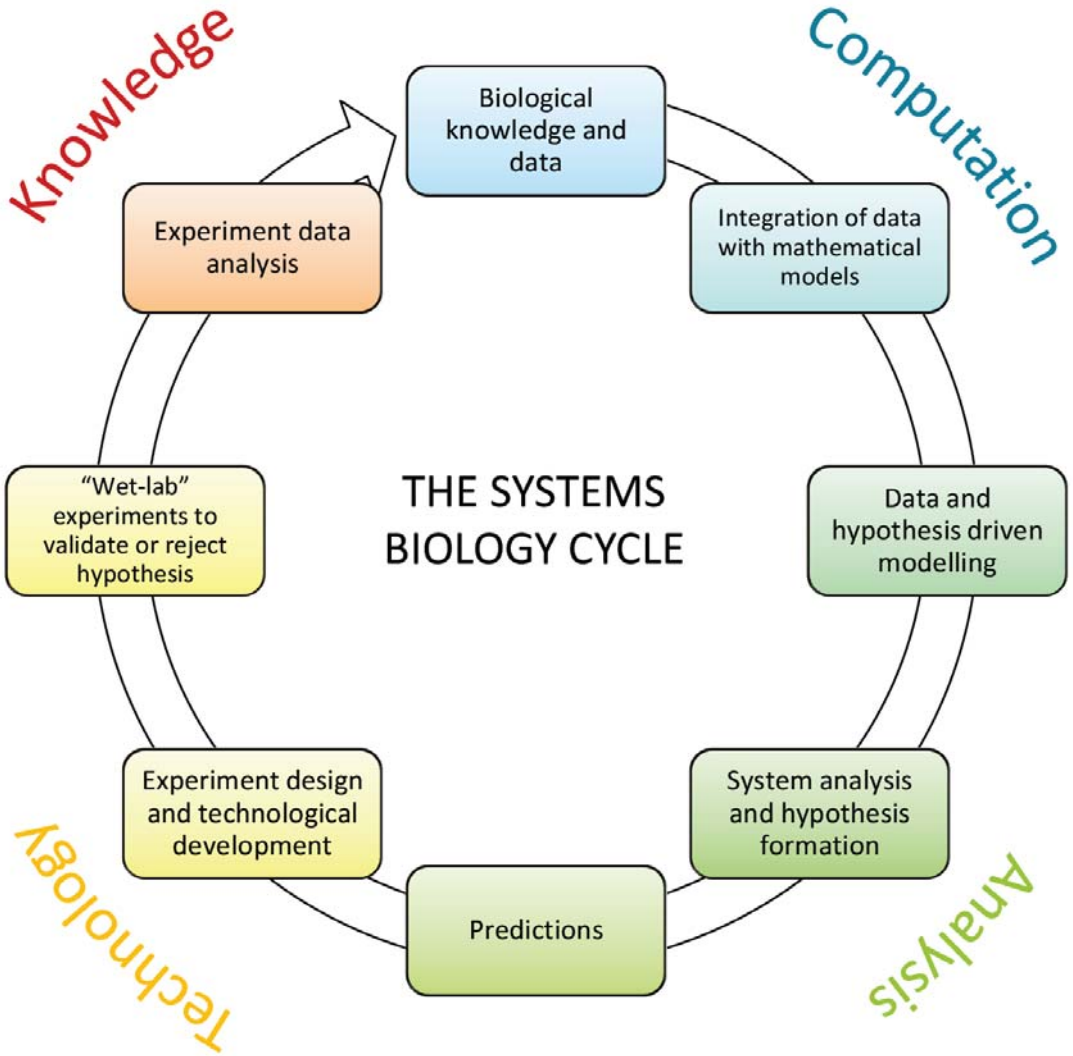


Figure 2.1: The systems biology cycle.

2.4 Metabolic Networks

Metabolic networks are the study of the complete set of metabolic and physical processes, that are determined the physiological and biochemical properties of a cell and these networks are comprised the chemical reactions of metabolism, the metabolic pathways, as well as the regulatory interactions that guide these reactions. Metabolic network maps are represented by a complex chain of chemical reactions and are highly associated between genes, proteins and enzymes, which are controlling the fundamental mechanisms that govern the biological system. Metabolic and regulatory networks are often large and complex, the construction and analysis of their computational models can be useful to identify the structural properties of a metabolic network that's being links the cellular phenotype to the corresponding genotype. The phenotype of a biological cell can be studied by means of exploring cellular biochemical reactions.

2.4.1 Glossary

Metabolism: A small chemical compounds, known as metabolites, can be ingested and secreted across the cellular membrane by means of transport reactions. Metabolites are the organic compounds that are used in, or created by, the chemical reactions happening in every cell of living organisms. This process, known as metabolism, which are responsible for breaking down food and other chemicals into energy and materials required for health, growth, and reproduction. Metabolism is also responsible for the removal of toxic substances from the body. Metabolites can be the starting materials, intermediate materials, or end products of these chemical reactions. External metabolites are considered to have defended concentrations while internal metabolites have to fulfil a balance condition at steady state.

Substrate and Product: A molecule called a substrate enters a metabolic pathway depending on the needs of the cell and the availability of the substrate. An increase in

concentration of anabolic and catabolic intermediates and/or end-products may influence the metabolic rate for that particular pathway. Each single reaction converts one group of substrate metabolites into another group of product metabolites and catalyzed by a specific enzyme (Schilling et al., 1999; Lacroix et al., 2008; de Figueiredo et al., 2009). The connections between biochemical reactions through the substrate and product metabolites are created complex metabolic networks that may be analyzed using network theory, stoichiometric analysis, and information on protein structure or function and metabolite properties (Hatzimanikatis et al., 2004).

Flux: Each biochemical reaction is controlled or catalyzed by one or more enzymes and is characterized by its speed of execution, known as a reaction flux. A reaction in which the rate of the forward reaction is always so much higher than the rate of the reverse reaction that the latter is relatively negligible. A reaction with metabolites lying on the opposite side of a membrane of the cell is considered as an external reaction. Otherwise, a reaction with all of its metabolites exclusively lying within the cell boundaries is considered to be an internal reaction. The metabolite is considered internal to the network if it is found inside the cell, otherwise it is external in which case it is a substrate or product of an external reaction. Reaction with metabolites on the opposite side of the cellular or organelle membrane is informed as the transport reaction. It is important to note that transport reactions are the superset of the external reactions.

Reaction rate: The reaction rate of a certain reaction when the metabolic network is in the quasi steady - state.

Steady-state: The state in which the concentrations of every metabolite does not change.

Stoichiometric matrix: Matrix containing the stoichiometric constraints for every reaction in terms of each chemical reaction. Every row of this matrix represents one unique metabolite for a network with *m* compounds and every column corresponds to one reaction of total *n*, as shown in equation 2.1. The entries in each column are the stoichiometric coefficients of the metabolites participating in a reaction. If a metabolite is formed (produced) by the reaction, the coefficient has assigned a positive (+) sign, on the other side, if it is consumed by the reaction; the stoichiometric coefficient appears with a negative (-) sign. All other rows are zero, which means that the corresponding to metabolites that do not participate in the reaction networks. Usually, the stoichiometric matrix as denoted by **S** and defined as follows:

$$\begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1n_1} \\ S_{21} & S_{22} & \cdots & S_{1n_2} \\ \vdots & \vdots & \ddots & \vdots \\ S_{m_1 1} & S_{m_2 2} & \cdots & S_{mn} \end{bmatrix} \tag{2.1}$$

A simple example of the metabolic network as shown in figure 2.2, where A is the starting molecule, B and C are the intermediate metabolites and D is the product, i.e. the final output. Reaction 1, reaction 2 and reaction 3 are controlled or catalyzed by the enzyme 1, enzyme 2 and enzyme 3, respectively.

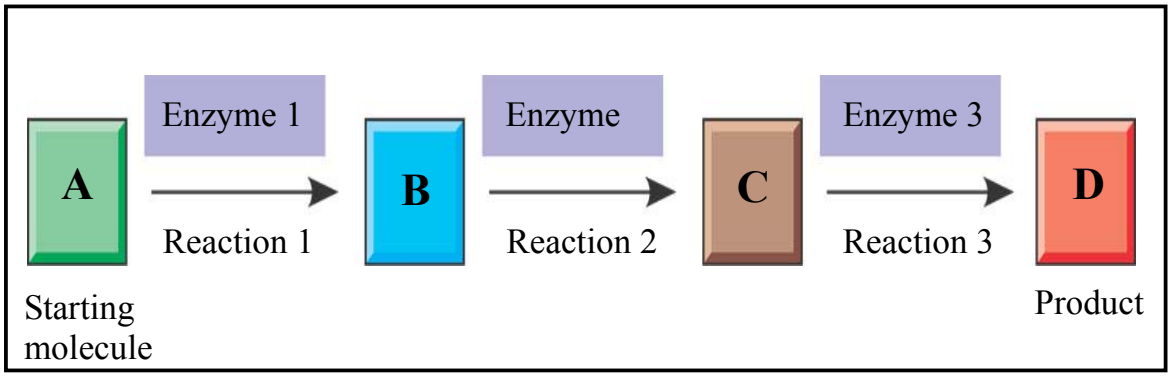


Figure 2.2: A Simple example of metabolic network

2.4.2 Mathematical Representation

The $m \times n$ stoichiometry matrix, $\mathbf{S}_{m \times n}$, is used to quantitatively represent the metabolic network with *m* internal metabolites and *n* reactions (Equation 2.1). The element in *l*-th row and *i*-th column of **S** represents the amount in moles of the *l*-th metabolite consumed or produced by the *i*-th reaction. The flux values for all the *n* reactions in the metabolic network are collected into a metabolic flux vector denoted here as $v_{n \times 1}$. Metabolic network reactions may be reversible or irreversible, where the flux of every irreversible reaction must be non-negative. Hence, an additional thermodynamic constraint must be imposed on the elements of the metabolic flux vector corresponding to the irreversible reactions as $v_i \geq 0$, where $i \in \text{irrev}$ are indices of the irreversible reactions. For the given stoichiometry network, the concentration of *m* metabolites and their change in time can be described using a system of ordinary differential equations as follows:

$$\frac{dC_l}{dt} = \sum_{i=1}^{n} S_{li} v_i \qquad \text{for } l = 1,...,m \tag{2.2}$$

Where, C_l denotes the concentration of the *l*-th metabolite in the network.

The pathway database provides the detailed information about the biochemical pathways for a given species of interest. Several databases are developed for the cellular enzymes and reactions in the public domain, such as KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa and Goto, 2012; Kanehisa et al., 2012), EcoCyc (Keseler et al., 2011), MetaCyc (Caspi et al., 2006 and 2008) and HumanCyc (Green et al., 2004). Metabolic network reconstructions are important and available for both prokaryotic and eukaryotic organisms such as Escherichia coli (Feist et al., 2009), Sacharomyces cerevisiae (Forster et al., 2003, Duarte et al., 2004), Haemophilus influenza (Schilling and Palsson, 2000), Helicobacter pylori (Schilling et al., 2002; Thiele et al., 2005), Mycoplasma genitalium

(Suthers et al., 2009), Staphylococcus aureus (Becker and Palsson, 2005), Homo sapiens (Duarte et al., 2007; Thiele et al., 2013) and so on.

2.4.3 Construction of Stoichiometric Modeling

The stoichiometric models are described a given metabolic network, which can be represented mathematically by a stoichiometric matrix, **S**. Construction of stoichiometric modeling is very important and essential process for further analysis of metabolic network model. In the construction of stoichiometric modeling, we consider a simple example of metabolic network model as shown in figure 2.3. The procedure for the construction of the stoichiometric model of a metabolic pathway network as follows:

Step-1: We obtain a set of pathways or reactions and their stoichiometry from pathway databases. Figure 2.3 shows a simple pathway map as an example of stoichiometric modeling. The extracellular metabolite substrate uptake A_{ex} into the system A with a flux v_1 is converted to B and C with flux v_2 and v_3, respectively, and metabolites B and C are subsequently excreted to product D and E at the respective flux rates v_4 and v_5. The reaction formulas are also listed in figure 2.3.

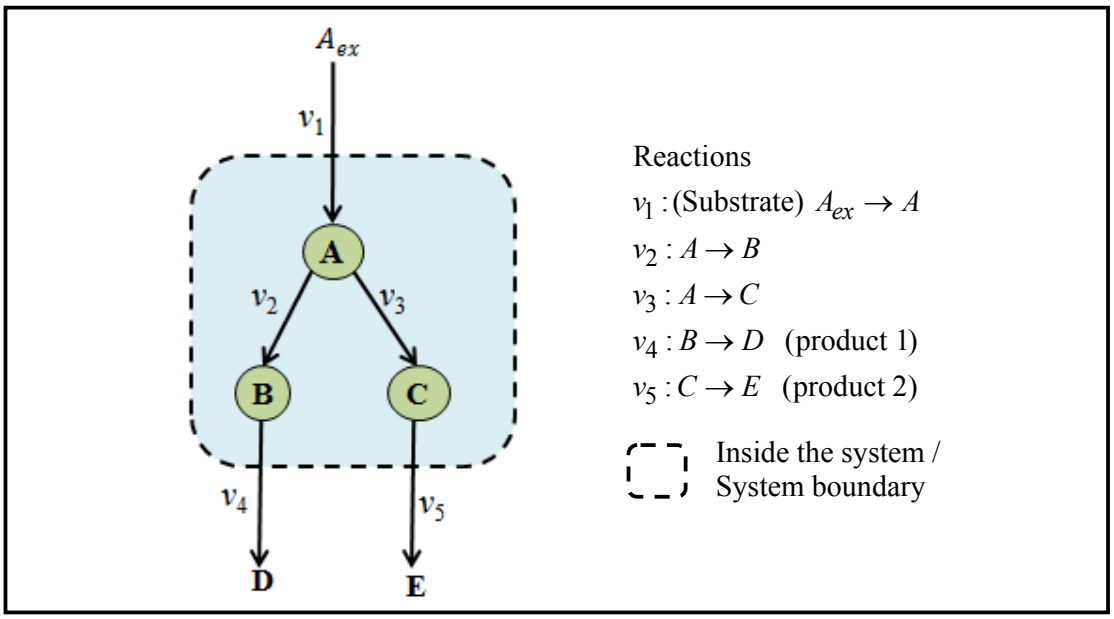


Figure 2.3: Example pathway of stoichiometric modeling. The dashed line represents the inside the system or system boundary, and nodes A, B, and C represent intermediate / intracellular metabolites and A_{ex}, D (product 1) and E (product 2) represent extracellular metabolites.

Step-2: Describe the mass balance equation for each intermediate metabolite in the pathway mathematically (e.g., A, B, C).

$$\frac{d[A]}{dt} = v_1 - v_2 - v_3 \tag{2.3}$$

$$\frac{d[B]}{dt} = v_2 - v_4 \tag{2.4}$$

$$\frac{d[C]}{dt} = v_3 - v_5 \tag{2.5}$$

Step-3: Use the steady-state assumption that the time-derivatives of these intermediate metabolite concentrations are zero i.e.,

$$\frac{d[A]}{dt} = \frac{d[B]}{dt} = \frac{d[C]}{dt} = 0 \tag{2.6}$$

Then, we can write the following equations from step-2, by using equation 2.6 as follows:

$$v_1 - v_2 - v_3 = 0 \tag{2.7}$$

$$v_2 - v_4 = 0 \tag{2.8}$$

$$v_3 - v_5 = 0 \tag{2.9}$$

Step-4: Thus, the mass balance equations can be described by a matrix form as follows:

$$\begin{array}{c} \\ d[A]/dt \\ d[B]/dt \\ d[A]/dt \end{array} \begin{array}{c} \begin{array}{ccccc} v_1 & v_2 & v_3 & v_4 & v_5 \end{array} \\ \begin{pmatrix} 1 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \end{pmatrix} \end{array} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix} = 0 \tag{2.10}$$

Stoichiometric matrix (S) Flux vector (v)

Finally, we can write the mass - balance equation mathematically as follows:

$$\mathbf{S} \cdot \mathbf{v} = 0 \tag{2.11}$$

2.5 Constraint-Based Metabolic Network Analysis

Constraint-based metabolic network analysis are the promising tools for the study of metabolic networks, as they do not require detailed knowledge of the biochemical reactions. Some of the methods only need information about the stoichiometric coefficients of the reactions and their reversibility types, i.e., constraints for steady-state conditions (Marashi et al., 2011). The traditional approach to metabolic modelling is to describe the components of a model in such detail that the model correctly represents the phenotype, then the constraint-based approach is rather to impose increasingly detailed constraints on the solution space so

that only relevant phenotypes are feasible. There are many applications and situations of interest, it may be assumed that the concentration of the metabolites internal to the metabolic network is constant in time (Clarke, 1988; Covert et al., 2003; Park et al., 2009). Which implies that the for every internal metabolite, the amount being produced is equal to the amount being consumed by the participating reactions. According to this condition, we can be imposed on the metabolic network as defined following two constraints as follows:

1. **Pseudo steady-state:** Internal metabolites are not accumulating within the metabolic network. In other words, the amount of metabolite being produced equals the amount being consumed.

$$\sum_{i=1}^{n} S_{li} v_i = 0 \qquad \text{for } l = 1, \ldots, m \tag{2.12}$$

2. **Feasibility:** $v_i \geq 0$, if $i \in irrev$ where irrev is a set of indices of irreversible reaction, i.e. the flux is non-negative for irreversible reaction.

These constraints are fundamental and used during the process of the reconstruction of the genome-scale metabolic network model. If the reconstructed metabolic network models are available, there may be two possible approaches for its analysis will be described in a next section, namely, (1) optimization-based and (2) pathway-based.

2.5.1 Optimization-based Analysis Methods

At quasi-steady state and the reaction thermodynamics, evolutionary nature and goals of the biological cell may be included when the constraints imposed on the metabolic network. Particularly the case for the biological cells of microorganisms, such as various bacteria and fungi, as well in some other cell types in eukaryotes which strive to optimize the biomass function, cellular growth, or in some cases energy production (ATP). The maximization of the cellular growth and its respective biomass reaction is an adaptive evolutionary nature of the biological cell. Biomass reaction is an artificial reaction added to

the metabolic network model and it assures that all of the metabolites necessary for the cellular growth are present in an experimentally determined proportion (Feist and Palsson, 2010). It is important to mention that while the cellular growth may be adopted as a valid optimization objective function to optimize its metabolic network flux distribution dictated by some other biological goal. Therefore, the metabolic network can be analyzed with an appropriately selected cellular objective and under the given constraints using various linear and non-linear optimization methods (Zomorrodi et al., 2012). To analyze the metabolic network and the cellular phenotype by the optimization-based methods can be used on either (1) wild-type or (2) mutant-type cell. Mutant cell is obtained using genetic modification techniques such as knock-out, knock-in or over-expression of the genes, which are responsible for the reactions in the metabolic network. In illustration of the methods, it will be assumed that the metabolic network is given with its stoichiometry matrix $\mathbf{S}_{m\times n}$, and the corresponding flux vector with $\mathbf{v} = v_{n\times 1}$. In the case when wild-type cell flux vector is contrasted with the mutant cell flux vector, the notation $v^{(wt)}$ and $v^{(mut)}$ will be used, respectively.

2.5.1.1 Flux Balance Analysis

Flux balance analysis (FBA) is an optimization-based approach that used to predict the quasi- or steady-state metabolic fluxes by applying mass balance constraints and objective functions (Fell and Small, 1986; Varma and Palsson 1994a,b; Kauffman et al., 2003; Lee et al., 2006; Feist and Palsson, 2010; Raman and Chandra, 2009; Orth et al., 2010). FBA is a constraint-based mathematical method that predicts the internal fluxes of large-scale metabolic networks, without requiring the biochemical knowledge of the network such as concentration of metabolites or enzyme kinetics of the system that make it easy to implement. Maximization of the biomass function or ATP production is frequently used as an objective

function for predicting the metabolic fluxes in the exponential growth phase (Van and Heijnen, 1995). We can evaluate the maximum yield of the specific compound using maximization of the target production rate as the objective function. Because FBA can be performed from the network information alone and without the enzyme kinetics, many FBA studies use genome-scale metabolic pathways rather than a small pathway alone, such as the central carbon metabolism pathway. A stoichiometric modeling is first constructed, as in the general form of MFA (details in section 2.5.2.1), to predict the metabolic flux distributions using FBA. FBA is particularly useful when the metabolic network system is under-determined (details of the degrees of freedom and system are in section 2.5.2.1). The solution space of an under-determined system is then limited by the addition of constraints, such as the upper or lower limits of each flux, and a unique flux distribution is then predicted by applying an objective function.

In FBA, generally a linear objective function is used, the metabolic flux distribution that maximizes or minimizes the objective value can be solved by the following linear programming (LP) as follows.

$$\begin{aligned} &\text{maximize } Z = \sum_{i=1}^{n} c_i^T \cdot v_i \\ &\text{subject to} \quad \sum_{i=1}^{n} \mathbf{S}_{li} \cdot v_i = 0 \quad ; \quad v_i^{lb} < v_i < v_i^{ub} \end{aligned} \tag{2.13}$$

Where, c_i is a weight coefficient for flux v_i and the superscripts lb and ub represent lower and upper limits, respectively. The overview of FBA is shown in figure 2.4. Gene regulatory network information was integrated with the metabolic network to result in the methods such as regulatory FBA (rFBA) (Covert et al., 2001; Covert and Palsson, 2002, 2003), steady-state regulatory FBA (SR-FBA) (Shlomi et al., 2005).

Although FBA is an influential method of solving under-determined systems, but the choice of an appropriate objective function can be biased and requires careful consideration.

Thus the objective function are not the best choice. While the maximization of biomass yield is frequently used as the objective function in many studies, it is not certain whether a single objective function can be universally applicable, especially for gene knockout mutants (Toya et al., 2010). To solve this problem, advanced methods have been proposed such as MOMA (Segre et al., 2002), ROOM (Shlomi et al., 2005), FVA (Mahadevan and Schilling, 2003), OptKnock (Burgard et al., 2003), RobustKnock (Tepper and Shlomi, 2009), OptReg (Pharkya and Maranas, 2006), OptORF (Kim and Reed, 2010), OptForce (Ranganathan et al., 2010).

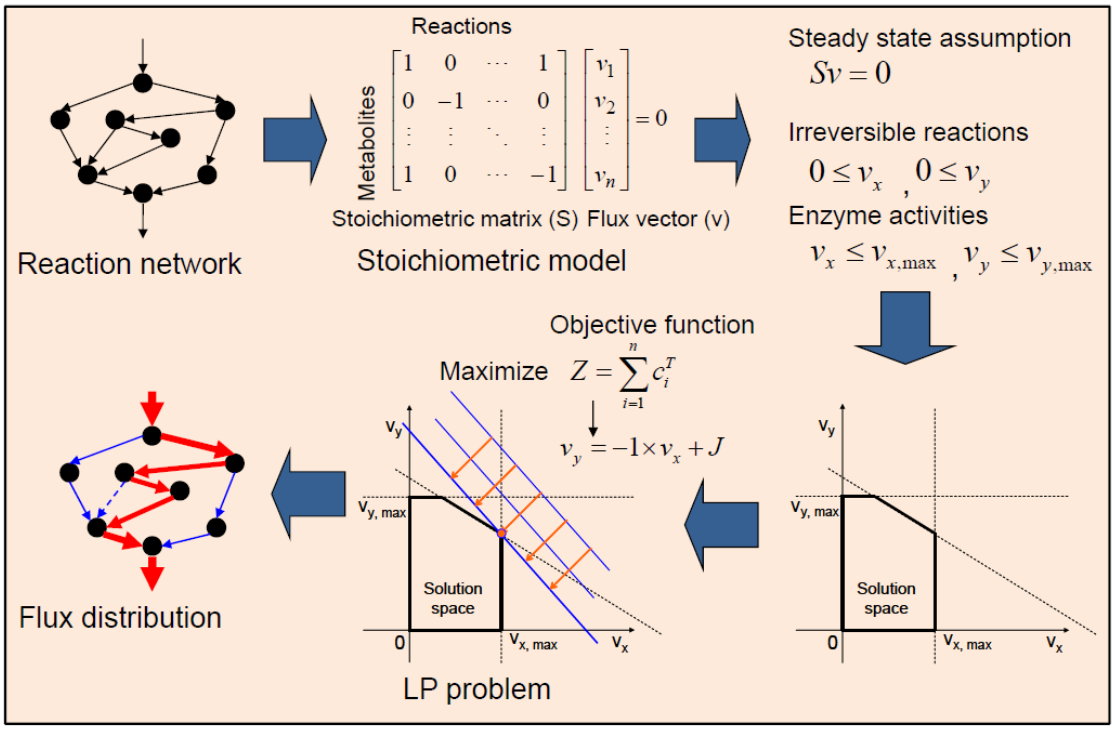


Figure 2.4: Overview of flux balance analysis (FBA)

2.5.1.2 Regulatory-Flux Balance Analysis

FBA has been used successfully to predict the time course of growth and by-product secretion, the effects of mutation and knockouts, and gene expression profiles. However, FBA leads to incorrect predictions in situations where regulatory effects are a dominant

influence on the behavior of the organism. FBA has not accounted for the constraints associated with regulation of gene expression nor activity of the expressed gene product. Therefore, regulatory-FBA (rFBA) has been proposed to include regulatory events within FBA to broaden its scope and predictive capabilities (Covert et al., 2001). Where, the transcriptional regulatory events represent as time-dependent constraints on the capabilities of a reconstructed metabolic network to further constrain the space of possible network function. Information of gene expression is incorporated by the Boolean logic formalism that uses a binary system, where the flux of one reaction is set to be zero if the relative gene is not expressed. The flux distribution of such gene knockout mutants could be optimized by LP under the additional constraint.

2.5.1.3 Minimization Of Metabolic Adjustment

Minimization of metabolic adjustment (MOMA) employs the quadratic programming (QP) to recognize a point in metabolic flux space, which is closest to the wild-type cell, compatible with the gene deletion constraint (Segre et al., 2002). The mutant cell would strive to minimize its overall flux distribution $v^{(mut)}$ deviation from the flux distribution in the wild-type cell $v^{(wt)}$, when a subset of reaction knockouts was performed in a cell. In MOMA, the flux distributions of gene knockout mutants can be estimated by the QP-based minimization of the Euclidian distance from those of wild type to those of a mutant as following problem:

$$\begin{aligned} &\text{minimize } Z = \sum_{i=1}^{n}\left(v^{(wt)} - v^{(mut)}\right)^2 \\ &\text{subject to} \quad \mathbf{S}\cdot v^{(mut)} = 0 \\ &\qquad\qquad v_h^{(mut)} = 0, \quad h \in KO \quad \text{where } KO \text{ is a set of indices of deleted reactions.} \end{aligned} \tag{2.14}$$

2.5.1.4 Regulatory On/Off Minimization

Regulatory on/off minimization (ROOM) uses mixed integer linear programming (MILP) to predict the metabolic flux distributions of gene deletion mutants in which the number of significant flux changes is minimized compared with wild type (Shlomi et al., 2005). The ROOM is motivated by two assumptions that (1) genetic regulatory changes required by flux changes after the reactions are knocked out are minimized by the cell in order to minimize the adaptation cost and (2) regulatory changes can be described using Boolean on/off dynamics which assigns fixed cost to each regulatory change irrespective of its magnitude. The MILP problem as follows:

$$
\begin{aligned}
&\text{minimize } Z = \sum_{i=1}^{n} y_i \\
&\text{subject to} \quad \mathbf{S} \cdot v^{(mut)} = 0 \\
&\qquad v_{\min,i} \le v^{(mut)} \le v_{\max,i} \\
&\qquad v_h^{(mut)} = 0, \quad h \in KO \\
&\qquad v_i^{(mut)} - y_i \left(v_{\max,i} - v_{i,u}^{(wt)} \right) \le v_{i,u}^{(wt)} \\
&\qquad v_i^{(mut)} - y_i \left(v_{\min,i} - v_{i,l}^{(wt)} \right) \ge v_{i,l}^{(wt)} \\
&\qquad y_i \in \{0,1\} \\
&\qquad v_{i,u}^{(wt)} = v_i^{(wt)} + \gamma \left| v_i^{(wt)} \right| + \zeta \\
&\qquad v_{i,u}^{(wt)} = v_i^{(wt)} - \gamma \left| v_i^{(wt)} \right| + \zeta \\
&\qquad i \in \{1,\ldots,n\} \setminus KO
\end{aligned}
\tag{2.15}
$$

The flux vectors $v^{(wt)}$ and $v^{(mut)}$ are flux distributions of the wild-type and mutant metabolic networks, respectively. Flux interval $\left[v_l^{(wt)}; v_u^{(wt)} \right]$ determines the local interval around the wild-type point. This local interval is determined using user specified parameters γ and ζ. Objective function, a sum of binary variables y_i, minimizes the number of unconstrained

reaction fluxes which can significantly deviate from their corresponding flux values in the wild-type metabolic network. If $y_i = 1$ which leads to no additional constraints on the flux $v_i^{(mut)}$. On the other side, if $y_i = 0$ the reaction $v_i^{(mut)}$ cannot significantly deviate from its respective wild-type value $v_i^{(wt)}$. One of the disadvantages of this method is the need to specify the parameters γ and ζ.

2.5.1.5 Flux Variability Analysis

Flux variability analysis (FVA) is often used to determine the robustness of metabolic models in various simulated conditions (Mahadevan and Schilling, 2003; Gudmundsson and Thiele, 2011). The optimal value of the given cellular objective (e.g. Biomass) often used in different flux distributions in the metabolic network analysis. If a maximal biomass flux is computed $v_{biomass}^{(\max)}$, it may then be appended as an equality constraint across multiple linear programs which all aim to determine a possible flux range of remaining reactions as follows:

$$\begin{aligned} &\max/\min && v_i && \text{for } i \in \{1,\ldots,n\} \\ &\text{subject to} && \mathbf{S}\cdot\mathbf{v} = 0 \\ & && v_{biomass} = v_{\max,biomass} \\ & && v_{\min,h} \le v_h \le v_{\max,h} && h \neq i \end{aligned} \tag{2.16}$$

2.5.1.6 OptKnock

The principal challenge in biotechnology is the systematic development of engineered microbial strains for optimizing the production of biochemical or chemicals (Stephanopoulos et al., 1998). However, the product yields of many microorganisms are often far below their theoretical maximums, when the absence of metabolic and genetic engineering interventions. OptKnock framework is developed a bi-level optimization for

suggesting gene knockout strategies for biochemical overproduction while recognizing that metabolic flux distributions are governed by internal cellular objectives (Burgard et al., 2003; Pharkya et al., 2003). The framework of OptKnock in the form of a bi-level MILP problem as follows:

$$
\begin{aligned}
&\max_{y} \quad v_{chemical}^{(mut)} \\
&\text{subject to} \quad \max_{v} \quad v_{biomass}^{(mut)} \\
&\qquad \text{subject to} \quad \mathbf{S} \cdot v^{(mut)} = 0 \\
&\qquad\qquad v_{biomass}^{(mut)} \geq \eta \, v_{\max,biomass}^{(wt)} \\
&\qquad\qquad v_{\min,i} \cdot y_i \leq v_i^{(mut)} \leq v_{\max,i} \cdot y_i \\
&\qquad y_i \in \{0,1\} \\
&\qquad \sum_{i=1}^{n} (1 - y_i) \leq K
\end{aligned}
\tag{2.17}
$$

Parameter η (usually =0.05-0.1) determines the required minimum flux which the biomass reaction should carry as a fraction of the maximum possible value $\left(v_{\max,biomass}^{(wt)} \right)$ in the wild type strain. The bi-level MILP problem (Equation 2.17) can be transformed into a one-level MILP and then solved using appropriate solver. The number of reaction deletions is constrained with the parameter *K*, where $y_i = 0$ denotes the inactivated reaction.

2.5.1.7 RobustKnock

A major disadvantage of the OptKnock framework was the existence of competing pathways with uncoupled production of the chemical with biomass. The OptKnock do not estimate for the presence of competing pathways in a metabolic network that may diverge metabolic flux away from producing a required chemical, resulting in lower or even zero chemical production rates in reality making these methods slightly over optimistic. This problem was addressed in a modified version of OptKnock, called RobustKnock, that

accounting for the presence of competing pathways in the network by predicting gene deletion strategies that lead to the over-production of chemicals of interest (Tepper and Shlomi, 2009). The framework of RobustKnock is formulated by a bi-level maximum-minimum optimization problem that underlies is the following form:

$$\begin{aligned} &\max_{y} \ \min_{v} \quad v_{chemical}^{(mut)} \\ &\text{subject to} \quad \max_{v} \quad v_{biomass}^{(mut)} \\ &\qquad \text{subject to} \quad \mathbf{S} \cdot v^{(mut)} = 0 \\ &\qquad\qquad v_{biomass}^{(mut)} \geq \eta \, v_{\max,biomass}^{(wt)} \\ &\qquad\qquad v_{\min,i} \cdot y_i \leq v_i^{(mut)} \leq v_{\max,i} \cdot y_i \\ &\qquad y_i \in \{0,1\} \\ &\qquad \sum_{i=1}^{n} (1 - y_i) \leq K \end{aligned} \tag{2.18}$$

Similarly as in OptKnock, the outer max-min problem searches for the reaction knockout subset of the size not larger than K, while the inner problem is the flux balance analysis which optimizes biomass flux for the given knockout combination.

2.5.1.8 OptReg

The OptReg framework was proposed to allow knockout, over-expression and under-expression of reactions in a metabolic network (Pharkya and Maranas, 2006). The allowed flux value range $\left[v_{\min,i}, v_{\max,i}\right]$ can be easily determined, but the interval $\left[v_{\min,i}^{(wt)}, v_{\max,i}^{(wt)}\right]$ corresponding to the wild-type flux value range requires experimental measurements. The above both interval are determined the optimization problem that underlies OptReg is formulated as follows:

$$
\begin{aligned}
&\max_{y_i^K, y_i^U, y_i^D} \quad \mathrm{v}_{chemical}^{(mut)} \\
&\text{subject to} \quad \max_{v} \quad \mathrm{v}_{biomass}^{(mut)} - \phi \cdot \sum_i v_i^{(mut)} \\
&\qquad \text{subject to} \quad \mathbf{S} \cdot \mathrm{v}^{(mut)} = 0 \\
&\qquad\qquad \mathrm{v}_{biomass}^{(mut)} \geq \eta \ \mathrm{v}_{\max,biomass}^{(wt)} \\
&\qquad\qquad \mathrm{v}_i^{(mut)} \leq \left[\mathrm{v}_{\min,i}^{(wt)} (1-C) + v_{\min,i} C \right] \cdot \left(1 - y_i^d\right) + v_{\max,i} \cdot y_i^d \\
&\qquad\qquad \mathrm{v}_i^{(mut)} \geq \left[\mathrm{v}_{\max,i}^{(wt)} (1-C) + v_{\max,i} C \right] \cdot \left(1 - y_i^u\right) + v_{\min,i} \cdot y_i^u \\
&\qquad\qquad \left(1 - y_i^k\right) + \left(1 - y_i^u\right) + \left(1 - y_i^d\right) \leq 1, \quad \forall \in \{1, \ldots, n\} \\
&\qquad\qquad \mathrm{y}_i^k \in \{0,1\}; \mathrm{y}_i^u \in \{0,1\}; \mathrm{y}_i^d \in \{0,1\}; \\
&\qquad\qquad \sum_i \left(1 - y_i^k\right) + \left(1 - y_i^u\right) + \left(1 - y_i^d\right) \leq K
\end{aligned}
\tag{2.19}
$$

Where, C is the regulation strength parameter in the interval [0; 1] and determines the fraction of the range to which down-regulated and up-regulated flux belongs. Parameter K is the maximal number of reactions which can be modified (deleted, up-regulated, down-regulated) and ϕ has a value determined through a trial-and-error process.

2.5.1.9 OptORF

The computational designs based on reaction deletions can sometimes result in strategies that are genetically complicated or infeasible, due to the presence of multi-functional enzymes and isozymes. Furthermore, due to regulatory restrictions the strains might not be able to grow initially. To address such aforementioned limitations, a new approach has developed for identifying metabolic engineering strategies based on gene deletion and over-expression, namely, OptORF (Kim and Reed, 2010). The OptORF is a bi-level optimization framework, which extends OptKnock to incorporate the GER (gene-

enzyme-reaction) associations. The framework of OptORF has outlined the following optimization problem as defined in equation 2.20.

$$
\begin{aligned}
\max \quad & v_{chemical}^{(mut)} - \alpha \sum_g z_g - \beta \sum_g w_g \\
\text{subject to} \quad & d_i \geq b_n \qquad \forall\ i \in J_{GER,n} \in S(j) \\
& d_i \leq \sum_{n \in S(j)} b_n, \qquad \forall\ i \in J_{GER} \\
& y_g \geq a_m, \qquad \forall\ g \in G_{MET} \cup G_{TF} \\
& y_g \leq \sum_{m \in M(g)} a_m, \qquad \forall\ g \in G_{MET} \cup G_{TF} \\
& (1 - a_m) \leq \sum_{r \in R^{Act}(m)} (1 - x_r) + \sum_{r \in R^{\mathrm{Rep}}(m)} x_r, \forall\ m \in M \\
& a_m \leq x_r, \qquad \forall\ m \in M, r \in R^{Act}(m) \\
& a_m \leq (1 - x_r), \qquad \forall\ m \in M, r \in R^{\mathrm{Rep}}(m) \\
& z_g - w_g = y_g - y'_g \quad \forall\ g \in G_{MET} \\
& z_g = y_g - y'_g \quad \forall\ g \in G_{TF} \\
& z_g + w_g \leq 1 \quad \forall\ g \in G_{MET} \\
& \sum_g z_g \leq K_1 \ \text{ and } \ \sum_g w_g \leq K_2 \\
& (b_n - 1) \geq \sum_{g \in G(n)} (y'_g - 1), \qquad \forall\ n \in N \\
& b_n \leq y'_g, \qquad \forall\ m \in M,\ g \in G(n) \\
& x_g \leq y'_g, \qquad \forall\ g \in G_{TF}
\end{aligned}
\tag{2.20}
$$

$$
\begin{aligned}
\max \quad & v_{biomass}^{(mut)} \\
\text{subject to} \quad & \mathbf{S} \cdot v^{(mut)} = 0 \\
& v_{\min,i} \leq v_i^{(mut)} \leq v_{\max,i} \\
& v_i^{(mut)} = 0, \text{ if } d_i = 0, \ \forall\ i \in J_{GER} \\
& d_i, b_n, w_g, z_g, y_g, y'_g, a_m, x_r \in \{0,1\}
\end{aligned}
$$

where the GER is a three-dimensional array consists of binary variables representing the reaction (d_i), enzyme (b_n), and gene (y_g). These binary variables denote if gene g is expressed, so that the enzyme n is active inducing a non-zero flux in the reaction i. Denote

with J_{GER} a set of indices i of reactions which appear as triplets in *GER*, $N(i)$ be a set of indices n of enzymes, which are associated with reaction i in GER, and $G(n)$ be a set of indices of genes g which are associated with the enzyme n in *GER*. While accounting for GER associations, one may step in further to look at the metabolic and transcription factor genes with respective sets G_{MET} and G_{TF}. For the expression of the gene m, here indicated using the binary variable a_m, can be influenced by activator or repressor r, denoted here using the binary variable x_r. Sets of activators and repressors which may influence condition m for the expression of one or more genes are denoted $R^{Act}(m)$ and $R^{Rep}(m)$, respectively. Binary variables z_g and w_g denote if the gene g was deleted or over-expressed, allowing no more than K_1 deletions and K_2 over-expressions in the cell.

2.5.1.10 OptForce

The OptForce has been proposed to identify all the possible engineering interventions by categorizing reactions in the metabolic network model, which is depending upon whether their metabolic flux values are increase, decrease or become equivalent to zero to meet a pre-specified overproduction target (Ranganathan et al., 2010). The framework of OptForce mainly follows two steps as given below:

Step-1: Identifying flux range of reactions for the wild-type strain. Solving 2q linear Problems as follows:

$$\begin{aligned} &\text{max/min} && v_i && \text{for } i \in \{1,...,n\} \\ &\text{subject to} && \mathbf{S}\cdot\mathbf{v} = 0 \\ & && v_{biomass} \geq \psi v_{\max,biomass} \\ & && v_{\min,i} \leq v_i \leq v_{\max,i} && \forall i \in \{1,...,n\} \setminus biomass \\ & && v_{\min,i}^{(\exp)} \leq v_i \leq v_{\max,i}^{(\exp)} && \forall i \in E \end{aligned} \tag{2.21}$$

It is assumed that it was possible to obtain an experimental flux range for the subset of reactions indexed by elements of the subset in E

Step-2: Identifying flux range of reactions for the over-producing (mutant) strain. Similarly as in previous step 2q linear problems as follows:

$$\begin{aligned} &\text{max/min} && v_i && \text{for } i \in \{1,...,n\} \\ &\text{subject to} && \mathbf{S}\cdot\mathbf{v} = 0 \\ & && v_{biomass} \geq \psi v_{biomass}^{(\max)} \\ & && v_{chem} \geq v_{chem}^{(\max)} \\ & && v_{\min,i} \leq v_i \leq v_{\max,i} \quad \forall i \in \{1,...,n\} \setminus biomass \end{aligned} \tag{2.22}$$

2.5.2 Pathway-based Analysis Methods

Metabolic pathways are the complex chain of biochemical reactions that occurring within a cell and increasingly promising in evaluating inherent network properties in the biochemical reaction network model. Metabolic pathway analysis becomes a core method for constructing a mathematical model that predicts the metabolic flux distribution for large-scale metabolic networks. Pathway-based analysis considers only the two constraints given in the section 2.5, without the specifying the cellular objective function that used in the optimization-based methods. Pathway-based analysis emerges for constructing a mathematical or computational model that accesses of functional and structural the properties of metabolic networks. Pathway-based methods are capable of characterizing the entire solution space of the possible metabolic network states without imposing the cellular objective bias. Optimization-based methods are guided by the cellular objective, and capable of exploring only a portion of the entire solution space. Metabolic pathway solution space can be confined using an algorithm for the enumeration of extreme rays in the bounded polyhedron where the bounded polyhedron corresponds to the degenerate polytope as will be

illustrated in more details in the upcoming subsection. Stoichiometries modeling is an essential process for the study of metabolic networks by the pathway-based analysis method.

2.5.2.1 Metabolic Flux Analysis

Metabolic flux analysis (MFA) is a powerful and essential tool for the determination of metabolic pathway fluxes. In this method, the intracellular metabolic fluxes are calculated by using a stoichiometric model for the major intracellular reactions and applying a mass balance condition around the intracellular metabolites. The implementation of the MFA, a set of measured metabolic fluxes are required in addition to the stoichiometric model. Usually, the input and output fluxes of the metabolic network are measured as specific rates by the experimental study. One of the strongest advantages of the MFA over other simulation methods is that, the stoichiometric model and the input and output fluxes are the only requirement for the input of MFA (Toya et al., 2011). We describe the processes used for the calculation of MFA using the simple example pathway as shown in figure 2.2. The mass balance equation for intracellular metabolites A, B and C can be written as the following matrix form as follows:

$$\frac{d}{dt}\begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix} = 0 \qquad \Rightarrow \mathbf{S} \cdot v = 0 \tag{2.23}$$

Where, **S** is the stoichiometric matrix and ***v*** is the reaction vector or flux vector. In the case, v_1 and v_4 are measurable (known) in the intracellular metabolites A, B and C, the stoichiometric matrix can be separated into known and unknown parts as follows:

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_4 \end{bmatrix} + \begin{bmatrix} -1 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} v_2 \\ v_3 \\ v_5 \end{bmatrix} = 0 \tag{2.24}$$

Multiplying by the inverse of the unknown part of the stoichiometric matrix on both sides of the equation and moving the known part to the right side of the equation, the unknown fluxes (v_2, v_3 and v_5) can be expressed as a function of the measurable (known) fluxes (v_1 and v_4) as follows:

$$\begin{bmatrix} v_2 \\ v_3 \\ v5 \end{bmatrix} = -\begin{bmatrix} -1 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 \\ 0 & -1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_4 \end{bmatrix} \tag{2.25}$$

In principle, the unknown flux vector can be solved using the stoichiometry and the known measurable fluxes. We need to consider the degrees of freedom of the network before actually solving the problem for the fluxes, (Stephanopoulos et al., 1998). The degrees of freedom are the number of independent fluxes and is calculated as the following equation as follows:

$$d = n - k - nd \tag{2.26}$$

Where, d is the degrees of freedom, n is the number of fluxes, k is the number of constraints, and nd is the number of measurable or determined fluxes. The system is called a "determined system" if the number of degrees of freedom is 0, then the fluxes are determined as a unique solution by the intersection of the lines which represent constraints. Moreover, the system is an "over-determined system" if the number of degrees of freedom is less than 0, then the minimum norm and least-squares solution can be calculated using the Moore-Penrose pseudo inverse method (Penrose, 1955 and Stephanopoulos et al., 1998). Conversely, the system is an "under-determined system" if the number of degrees of freedom is greater than 0, then immense solutions exist because of the lack of constraints. The relationship between the

network system and the number of degrees of freedom as shown in figure 2.5. We can determine the unknown metabolic fluxes are using MFA quite easily for determining or over-determined system. Under-determined systems require more constraints to reach a particular restricted solution. The primary challenge in the use of MFA is that most of the biological networks are under-determined systems.

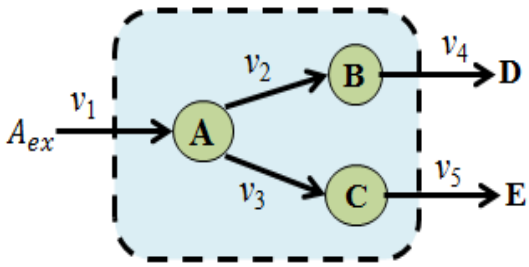
Degrees of freedom is defined as:

$d = n - k - nd$

n : The number of fluxes

k: The number of constraints

nd: The number of measurablefluxes



Example network

Known fluxes	v_1=100 v_4 = 40	v_1=100 ,v_4 = 40 v_5 = 30	v_1 = 100
Degrees of freedom	d=0 (n=5, k=3, m=2)	d=-1 (n=5, k=3, m=3)	d=1 (n=5, k=3, m=1)
System	Determined (d = 0)	Over-determined (d < 0)	Under-determined (d > 0)

Figure 2.5. Relationship between the number of degrees of freedom and the system.

2.5.2.2 Elementary Mode Analysis

Elementary mode (EM) analysis (Schuster and Hilgetag 1994) is one of the most popular and essential techniques in a metabolic pathway analysis of metabolic networks. The EM analysis is potentially effective for integrating transcriptome or proteome data into metabolic network and which is exploring the mechanism of how phenotypic or metabolic flux distributions are changed with respect to environmental and genetic perturbations (Zhao and Kurata, 2010). A quantitative measure of metabolic fluxes is carried by individual EMs,

which is of great opportunity to identify dominant metabolic processes, and to understand how these processes are redistributed in biological cells in response to the changes in environmental conditions, enzyme kinetics, or chemical concentrations.

Generally, the biological networks can be represented by a stoichiometric matrix **S**, for the interconnectivity of metabolites within a network of biochemical reactions. The rows of **S** represent to the metabolites (*m*) and the columns of **S** represent to the reactions (*n*) in a network, with elements corresponding to stoichiometric coefficients of the associated reactions. The coefficient of **S** has a positive (+) sign, if a metabolite is formed (produced) by the reaction and the stoichiometric coefficient appears with a negative (-) sign, if it is consumed by the reaction. All other rows are zero if the corresponding metabolites that do not participate in the reaction. At the steady-state, the mass balance equation is given by the following form as follows:

$$\mathbf{S} \cdot \mathbf{v} = 0, \tag{2.27}$$

where $\mathbf{v} = (v_1, v_2, ..., v_n)^t$ is the vector whose elements corresponds to metabolic fluxes and *n* is the number of reactions. The sets of basis vectors are determined by the all possible solution set of the equation (2.27). The EM matrix **P** is uniquely determined using the stoichiometric matrix and the flux vector, as follows:

$$\mathbf{v} = \mathbf{P} \cdot \boldsymbol{\lambda}, \tag{2.28}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_{ne})^t$ is the EMC vector and *ne* is the number of EMs. The components of these vectors and matrix are defined as the following form as follows:

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1m} \\ e_{21} & e_{22} & \cdots & e_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nne} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{ne} \end{pmatrix} \tag{2.29}$$

The j-th column for the **P** matrix is the j-th EM vector: $e_j = (e_{1j}, e_{2j}, ..., e_{nj})^t$. The flux distribution can be also represented as superposition of the EM vectors with non-negative EMCs as follows:

$$\mathbf{v} = \sum_{j=1}^{ne} \boldsymbol{\lambda}_j \mathbf{e}_j \tag{2.30}$$

The EMs fulfills the following three conditions as given below:

1. **Pseudo steady-state**: $\mathbf{S} \cdot \mathbf{v} = 0$, i.e. The intermediate or intracellular metabolite concentrations remain constant in the metabolic network model.
2. **Feasibility / Thermodynamics**: $\mathbf{v}_i \geq 0$, if the i-th reaction is irreversible.
3. **Non-decomposability / elementary**: There is no other vector like, **w** (**w** $\neq$ **v** and **w** $\neq$ 0), which is fulfilling the conditions 1 and 2, such that the set of indices of the non-zero elements in **w** is a strictly proper subset of the set of indices of the non-zero elements in **v**.

The Expa or cEM analyses were performed in the same manner as EM analysis, where the Expas or cEMs are employed instead of the EMs in equations (2.28, 2.29, 2.30). Details of the cEM analysis will be described in the materials and method section.

The EM analysis is a powerful metabolic pathway analysis tool to recognize the structure of a metabolic network. The EM analysis can decompose the complicated metabolic network by comprising of highly interconnected reactions into uniquely organized pathways. These pathways, consisting of a minimal set of enzymes that can operate at the steady-state condition of the cellular metabolism, which represent the independent cellular physiological states (Trinh et al., 2009). However, EM analysis does not decompose the reversible reactions into two irreversible reactions in calculating EMs and introduces a systematic way of

extracting biologically meaningful pathways from an intricate metabolic network. In this context, an alternative approach has been proposed, called, extreme pathway analysis (Expa) (Schilling et al., 2000).

2.5.2.3 Extreme Pathways Analysis

The extreme pathway (Expa) analysis is closely related to EM analysis because Expas are a subset of EMs. The Expa analysis can be measured as a mixture between stoichiometric model and EM analysis. The only internal reversible reaction is split into two irreversible reactions for computing of Expas, while it does not decompose reversible exchange reaction (Trinh et al., 2009). Different from EM analysis, Expa analysis contains two additional constraints, one of them to make all Expas systematically independent (Schilling et al., 2000). The metabolic flux vector can be expressed as a nonnegative linear combination of Expas or EMs in metabolic reaction networks. The Expas are the systematically independent subset of EMs; that is, one Expas can not be expressed as a nonnegative linear combination of any other Expas. The two additional conditions for Expa with EM as follows:

4. **Network reconfiguration**: Define exchange and internal reaction, all reversible reactions split into pairs (forward and backward direction) of irreversible reactions.

5. **Systemic independence**: The Expas are the systematically independent subset of EMs; that is, no Expas can be expressed as a nonnegative linear combination of any other Expas.

Note that, if the reactions, including both internal and exchange reactions are irreversible in metabolic networks, then the two sets EMs and Expas are identical or equivalent. On the other hand, if the exchange fluxes are all reversible, there is more EMs than Expas. Therefore, the identification of Expas depends on the metabolic network reconfiguration, while the

identification of EMs does not. For illustration, Expas identified in a metabolic network whose reversible exchange reactions split into two irreversible reactions may not be Expas to any further extent in the original metabolic networks whose reversible exchange reactions do not split (Klamt and Stelling 2003). The basic difference between EM and Expa analyses are illustrated in figure 2.6.

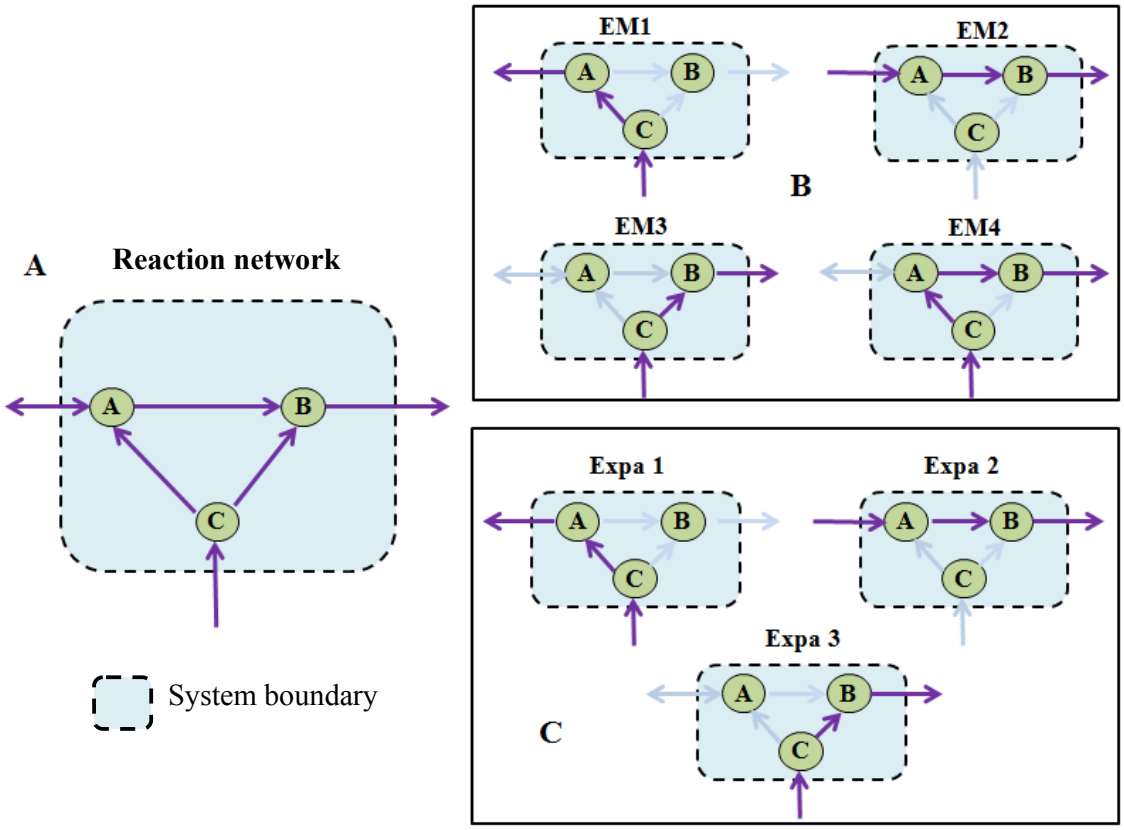


Figure 2.6: Difference between EM and Expa analyses. (A) Reaction network, (B) EM and (C) Expa analyses.

2.5.2.4 Control Effective Flux

The control effective flux (CEF) is an EM-based algorithm that has been developed to predict the transcriptional regulations, e.g. the gene expression patterns of *E. coli* (Stelling et al., 2002) and *Saccharomyces cerevisiae* (Cakir et al., 2004, 2007) grown on different

substrates. The CEF algorithm was developed to estimate the change in transcriptional regulations based on the topology of metabolic networks with specific biological reactions, when the substrate changes, e.g. from glucose to acetate, ethanol or glycerol (Cakir et al., 2004, 2007; Stelling et al., 2002). The efficiency of an EM corresponding to the specific biological function is denoted by, $\gamma_{j,\mathrm{SPEOBJ}}$. The ratio of the EM output those reactions are involving the objective function to the investment required to form each EM, i.e. the sum of the absolute elements in the EM is defined as follows:

$$\gamma_{j,\mathrm{SPEOBJ}} = \frac{P_{\mathrm{SPEOBJ},j}}{\sum_i \left| P_{i,j} \right|}, \tag{2.31}$$

where $P_{i,j}$ is the *i*-th reaction in the *j*-th EM of the normalized element and SPEOBJ is the related reaction number for the specific biological function, e.g., biomass production and ATP yield. The CEF is represented by the weighted sum of the *i*-th elements of the all EMs using their associated efficiency $\gamma_{j,\mathrm{SPEOBJ}}$ for the *i*-th reaction as follows:

$$\mathrm{CEF}_i = \sum_{\mathrm{SPEOBJ}} \frac{1}{P_{\mathrm{SPEOBJ}}^{\max}} \frac{\sum_j \left(\gamma_{j,\mathrm{SPEOBJ}} \cdot \left| P_{i,j} \right| \right)}{\sum_j \gamma_{j,\mathrm{SPEOBJ}}}, \tag{2.32}$$

where $P_{\mathrm{SPEOBJ}}^{\max}$ is the maximum element in the row of biological function. To assess the validity of the CEF metrics, the transcription ratio for the *i*-th reaction under different substrate conditions, S_1 and S_2, is defined by:

$$\theta_i \left(S_1, S_2 \right) = \frac{\mathrm{CEF}_i \left(S_2 \right)}{\mathrm{CEF}_i \left(S_1 \right)} \tag{2.33}$$

2.5.2.5 Modified Control Effective Flux

The modified CEF algorithm (mCEF) is necessary to apply CEF to a broad range of genetic mutants that over-express, under-express or lack a metabolic gene (Zaho and Kurata, 2009b). The efficiency of for such a genetic mutant for the *j*-th EM is defined by:

$$\gamma^{m}_{j,\mathrm{SPEOBJ}} = \frac{P_{\mathrm{SPEOBJ},j} \cdot \mathrm{EA}_j}{\sum_i \left(\left| P_{i,j} \right| \cdot \pi_i \right)}$$
$$\pi_i = \begin{cases} \mathrm{EAP}_i & \text{if reaction } i \text{ is modified} \\ 1 & \text{otherwise} \end{cases} \tag{2.34}$$

Where EAP_i is the enzyme activity parameter for the relative gene expression, i.e. enzyme activity, responsible for the *i*-th reaction of a mutant to wild type. When $\mathrm{EAP}_i = 0$, if the gene of the *i*-th reaction is deleted. If EAP*i* >1 or <1 then it is over- or underexpressed, respectively. The π_i , is the adjustment factor for calculating the investment for genetic mutants. EA_j is the adjustment factor that incorporates the change in the modified reaction into each EMs output, as defined by:

$$\mathrm{EA}_j = \prod_{i=1}^{n} \mathrm{ge}_{i,j}$$
$$\mathrm{ge}_{i,j} = \begin{cases} \mathrm{EPA}_i & \text{if } P_{i,j} \neq 0 \\ 1 & \text{if } P_{i,j} = 0 \end{cases} \tag{2.35}$$

Where, $\mathrm{ge}_{i,j}$ is the parameter demonstrating the gene expression state of the *i*-th reaction in the *j*-th EM. The numerator of the equation (2.34) is increasing or decreasing, if a gene within an EM is over-expressed or under-expressed respectively. For $\mathrm{EPA}_i = 0$, the EM containing it is neglected ($\gamma^{m}_{j,\mathrm{SPEOBJ}} = 0$), which is consistent with the EM analysis of gene

deletion mutants. When $\mathrm{EPA}_i = 1$, i.e. If gene expressions are not changed at all, then the equation (2.34) and equation (2.31) is consistent. Equation (2.34) is an extension of the original efficiency [Equation (2.31)] to genetic mutants. The mCEF for the mutant is defined as follows:

$$\mathrm{mCEF}_i\left(mut\right) = \sum_{\mathrm{SPEOBJ}} \frac{1}{P^{\max}_{\mathrm{SPEOBJ}}} \frac{\sum_j \left(\gamma^m_{j,\mathrm{SPEOBJ}} \cdot \left| P_{i,j} \right| \cdot \pi_i \right)}{\sum_j \gamma^m_{j,\mathrm{SPEOBJ}}} \tag{2.36}$$

The mCEF for wild type is the original of ECF as follows:

$$\mathrm{mCEF}_i\left(wt\right) = \sum_{\mathrm{SPEOBJ}} \frac{1}{P^{\max}_{\mathrm{SPEOBJ}}} \frac{\sum_j \left(\gamma_{j,\mathrm{SPEOBJ}} \cdot \left| P_{i,j} \right| \right)}{\sum_j \gamma_{j,\mathrm{SPEOBJ}}} \tag{2.37}$$

The mECF of the relative change in a gene expression profile of a mutant to wild type is provided by:

$$\theta_i\left(wt, mut\right) = \frac{\mathrm{mCEF}_i\left(mut\right)}{\mathrm{mCEF}_i\left(wt\right)} \tag{2.38}$$

2.5.2.6 Enzyme Control Flux

The EM-based algorithm, enzyme control flux (ECF) has been proposed (Kurata et al., 2007) to link enzyme activity data to flux distributions of metabolic networks that integrates enzyme activity into EM analysis. The ECF is describing how changes in enzyme activities between the wild-type and a mutant-type are related to changes in the EMCs by presents the power-law formula. The ECF have been validated by the integrated enzyme activity data into the EMCs of *E. coli* and *Bacillus subtilis* wild-type (Kurata et al., 2007). The ECF model

successfully uses an enzyme activity profile to estimate the flux distribution of the mutants and the increase in the number of incorporating enzyme activities decreases the model error of ECF (Kurata et al., 2007).

The EMCs of wild type $\lambda^{wt} = \left(\lambda_1^{wt}, \lambda_1^{wt}, ..., \lambda_m^{wt}\right)^t$ are calculated from the flux distribution of the wild type by the optimization problem of QP (Schwartz and Kanehisa, 2005, 2006) as follows:

$$\begin{aligned} \min \quad & \sum_j \left(\lambda_j^{wt}\right)^2 \\ \text{subject to } & v = \mathrm{P} \cdot \lambda^{wt} \\ & \lambda_j^{wt} \geq 0. \end{aligned} \tag{2.39}$$

Then the EMCs of a mutant are provided by the following form as follows:

$$\begin{aligned} \lambda_j^{mut} &= \beta \cdot \lambda_j^{wt} \prod_{i=1}^{n} a_{i,j} \\ a_{i,j} &= \begin{cases} a_i & \text{if } P_{i,j} \neq 0 \\ 1 & \text{if } P_{i,j} = 0 \end{cases} \end{aligned} \tag{2.40}$$

Where, EMC vector of a mutant type, $\lambda^{mut} = \left(\lambda_1^{mut}, \lambda_1^{mut}, ..., \lambda_m^{mut}\right)^t$, $a_{i,j}$ is the relative enzyme activity for the *i*-th reaction in the *j*-th EM of a mutant to wild type, a_i is the enzyme activity ratio for the *i*-th reaction of the mutant to wild type. β is the factor use for the normalized, so that the substrate uptake flux is the same as that of wild type. The flux distribution of the mutant type is provided by the following form as follows:

$$v^{mut} = P \cdot \lambda^{mut} \tag{2.41}$$

2.5.2.7 Genetic Modification of Flux

The critical technologies used for designing or improving the metabolic flux distribution of microbes by the gene deletion and over-expression. There are many algorithms already have been developed to predict a flux distribution from a stoichiometric matrix in the mutants in which some metabolic genes are deleted or non-functional, but there are few algorithms that predict how a broad range of genetic modifications, such as over- and under-expression of metabolic genes, alters the phenotypes of the mutants at the metabolic flux level. Genetic modification of flux (GMF) (Zhao and Kurata, 2009b) has been proposed to overcome such problem, which couples, two algorithms modified control effective flux (mCEF) and enzyme control flux (ECF). A flow chart of the GMF algorithm is depicted as shown in Figure 2.7. GMF is an EM-based algorithm that integrates gene expression or enzyme activity data to predict the flux distributions. The GMF algorithms not only used to predict the flux distribution of a gene deletion mutant, but also the mutants with under-expressed and overexpressed genes in *E. coli* and *Corynebacterium glutamicum* (Zhao and Kurata, 2009b).

The CEF ratios of a mutant to wild type are calculated from the metabolic network topology by the mCEF algorithm that are presented in earlier sections. Assume that enzyme activity profile is linearly correlated to its associated gene expression profile, the EMCs are calculated of a mutant cells from the flux distribution of the wild type optimization problem by QP [Equation (2.39)]. There is a quantitative correlation between mRNA expression and protein levels are found in a few cases (Ideker et al., 2001; Siddiquee et al., 2004). When the

ratios of enzyme activity can be replaced by the ratios of CEF, the EMCs for the mutant are provided by equation (2.40) as follows:

$$\lambda_j^{mut} = \beta \cdot \lambda_j^{w} \prod_{i=1}^{n} \theta_i \left(wt, mut \right) \tag{2.42}$$

Finally, when we get the EMCs of the mutant type, then the predicted flux distribution is provided by the following form as follows:

$$v^{mut} = \mathbf{P} \cdot \lambda^{mut} \tag{2.43}$$

2.6 Integrated Biological Network Analysis

The earlier cited and described methods are aim to incorporate heterogeneous biological data into metabolic network. Some of those which augment metabolic network with the gene expression, gene regulatory network or enzyme activity profile. The EM-based algorithms are potentially effective for integrating transcriptome or proteome data into metabolic network analyses and in exploring the mechanism of how phenotypic or metabolic flux distribution is changed with respect to environmental and genetic perturbations with a certain condition. In this section, we will present in more details about the recent effort of integrated biological network analysis aiming for more accurate prediction of the cellular phenotype for the given environment and growth conditions.

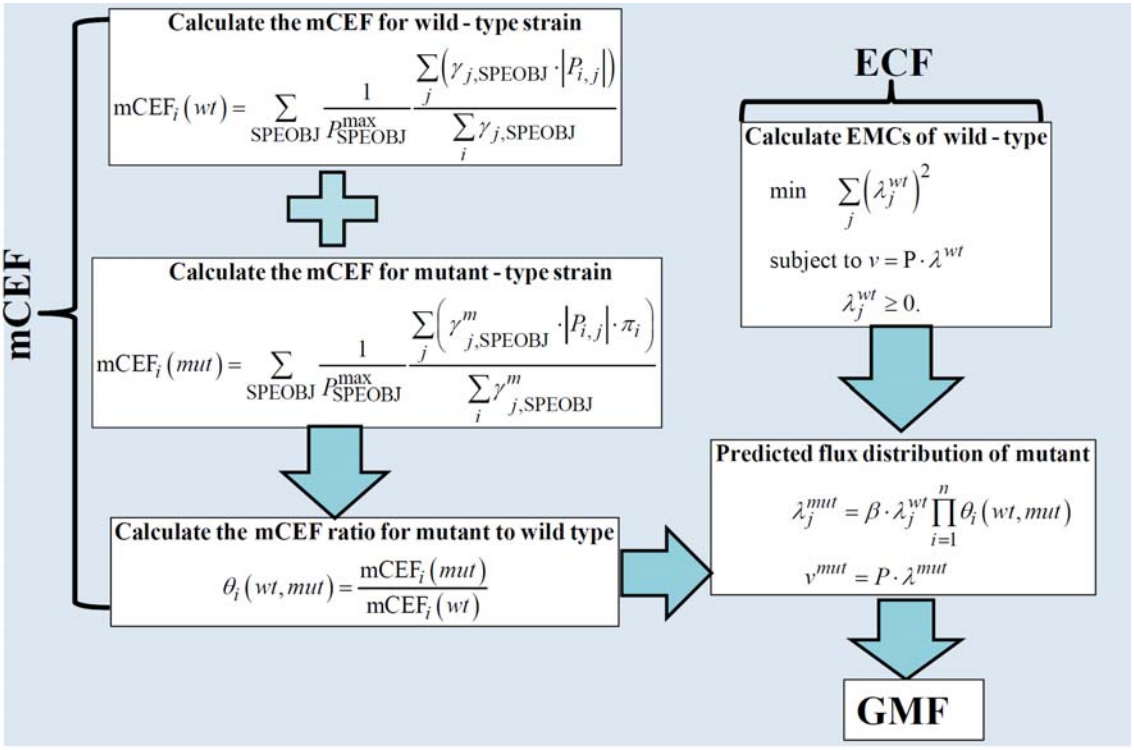


Figure 2.7: A flow chart of the GMF algorithm.

2.6.1 Integrative Omics-Metabolic Analysis

Integrative omics-metabolic analysis (IOMA) was proposed for the prediction of flux distribution from a metabolic network by integrating of quantitative proteomics and metabolomics data with a genome-scale metabolic models (Yizhak et al., 2010). IOMA was shown to successfully predict the metabolic state of human erythrocytes and *E. coli* under different gene knockouts (Yizhak et al., 2010). IOMA was displaying a significant improvement to compare with a less comprehensive method such as FBA, which depends on a definition of a cellular objective function or earlier described MOMA, which depends on data regarding the wild-type flux distribution. The general framework of IOMA formulated as a QP problem which minimizes the inconsistency between the flux predicted by the metabolic network model and the one obtained from the metabolomic and proteomic data using the Michaelis-Menten (MM) kinetics.

2.6.2 Integrative Metabolic Analysis Tool

Integrative Metabolic Analysis Tool (iMAT) have been proposed for scientifically predicting the human tissue-specific metabolic behavior by integrating the tissue-specific gene and protein expression data with the genome-scale metabolic network under the assumed steady-state constraints (Shlomi et al., 2008). The iMAT was done by modifying Boolean mapping values of 1 and 0 to account for expression states, highly and lowly expressed genes, respectively, and replacing with 'max' and 'min' by the logical operators 'and' and 'or' expressions, respectively. The expression data are given in two subsets R_H and R_L, which indicate the highly expressed (over- or under-expressed) and lowly expressed. The subset R_L used binary variable y_i^+, if the *i*-th reaction is truly lowly expressed in the problem. Similarly, R_H is used binary variable y_i^+ and y_i^-, if the *i*-th reaction is highly over-expressed and under-expressed, respectively. The framework of iMAT was formulated to find a steady-state flux distribution by satisfying the stoichiometric and thermodynamic constraints of the following MILP problem as follows:

$$
\begin{aligned}
\min \quad & \sum_{i \in R_H} y_i^+ + y_i^- + \sum_{i \in R_L} y_i^+ \\
\text{subject to} \quad & \mathbf{S} \cdot \mathbf{v} = 0 \\
& v_{\min,i} \leq\leq v_{\max,i} \quad ,\forall i \in \{1,...,n\} \\
& v_i + y_i^+ \left(v_{\min,i} - \varepsilon\right) \geq v_{\min,i} \quad ,\forall i \in R_H \\
& v_i + y_i^- \left(v_{\max,i} + \varepsilon\right) \leq v_{\max,i} \quad ,\forall i \in R_H \\
& v_{\min,i}\left(1 - y_i^+\right) \leq v_i \leq v_{\max,i}\left(1 + y_i^+\right) \quad ,\forall i \in R_L \\
& y_i^+ + y_i^- \leq 1 \quad , \forall i \in R_L \text{ and } y_i^+, y_i^- \in \{0,1\}
\end{aligned}
\tag{2.44}
$$

This integration of heterogeneous biological data is implemented in the iMAT tool as described elsewhere (Zur et al., 2010).

2.7 Network Redundancies and Inconsistencies

The cellular life is a highly redundant complex system and the evolutionary maintenance of the redundancy remains unexplained (Wang and Zhang, 2009). Metabolic network may have redundant metabolites and reactions, whereby eliminating them from the original network may significantly simplify the analysis. In the constraint-based metabolic network analysis of the metabolic networks, most of the known reduction methods are analogous to the methods for the reduction of linear equality and inequality constraints (Luenberger, 2003; Gagneur and Klamt, 2004). The known reduction techniques for network redundancies and inconsistencies as follows:

1. Metabolite conservation relations are a linear dependency related to the among rows of the stoichiometry matrix. The metabolic network can be simplified by removing the redundant metabolites having dependency relation in the network. It suffices to reduce the stoichiometry matrix **S** to a maximal linearly independent set of rows using simple linear algebra.
2. Strictly detailed, balanced reactions those values are zero in any quasi-steady state the network. Which can be identified by solving a simple linear program of the stoichiometry matrix **S**.
3. Uniquely produced (consumed) metabolite is a metabolite *m* which is produced (consumed) by a single reaction v_{t0}, and respectively consumed (produced) by *i-th* other reactions, namely $v_{t1}, v_{t2}, ..., v_{ti}$. Metabolite *m* and $i+1$ reactions can be removed and substituted with *i-th* new reactions.
4. Enzyme subset is defined as a subset of reactions at any steady state. If the enzyme subset has k reactions $v_{t1}, v_{t2}, ..., v_{ti}$, then $i-1$ of them can be eliminated from the network, as their later recovery is possible if a flux of the one remaining reaction is

known. The reactions of enzyme subsets are computed using the right null space matrix **R** of the stoichiometry matrix **S**.

2.8 Conclusion

The computational or mathematical models can be useful to identify the structure of a metabolic network that links the cellular phenotype to the corresponding genotype, although those networks are often large and complex. Constraint-based metabolic network analysis employs two approaches, optimization-based and pathway-based analysis, which are used for predicting the steady-state intracellular fluxes from the metabolic network by integrating of the experimental data from genomics, transcriptomics, proteomics, metabolomics, and fluxomics, which are determined by high-throughput technologies. Pathway-based analysis is the most widely used and more advantageous than optimization-based analysis, because it generally employs without the specifying the cellular objective function. Comparison between some optimization-based and pathway-based metabolic network analysis and its applications are shown in table 2.1. Metabolic pathway analysis aims to discover and analyze meaningful routes involved in the metabolic networks, linking the cellular behavior with its inherent metabolic network structure. In this chapter, we describe details about the integration of the heterogeneous biological data into the metabolic network of the optimization-based analysis methods, e.g., FBA, MOMA, ROOM, FVA, with some genetically modified algorithms, e.g., OptKnock, RobustKnock, OptReg, OptGene, OptGene, OptForce and pathway-based analysis methods, e.g., MFA, EM, Expa, CEF, mCEF, ECF, GMF. Also, we describe the integration of quantitative proteomics and metabolomics data, and tissue-specific gene and protein expression data with the genome-scale metabolic network by integrating biological network analysis.

Table 2.1: Comparison between some optimization-based and pathway-based metabolic network analysis and its applications based on (Jiang, 2006).

Approach	Constraints incorporated					Computational demand	Flux distributions	Application						
	Stoichiometry	Thermadynamics	Quasi steady state	Reaction capacities	Optimality			Functional pathways	Optimal operation	Importance of reactions	Correlated reactions	Pathway lengths	Network functionality	Robustness / Flexibility
MFA	√	√	√	√	×	Low	Single	×	×	√	×	×	×	×
FBA	√	√	√	√	×	Low	Single	×	√	√	×	×	√	√
MOMA	√	√	√	√	√	Medium	Single	×	√	√	×	×	√	√
EM	√	√	√	×	×	High	All	√	√	√	√	√	√	√
Expa	√	√	√	×	×	High	All	√	√	√	√	√	√	√

√ = Applicable , × = Not Applicable

Chapter 3

Materials and Methods

3.1 Introduction

We develop the complementary elementary mode (cEM) analysis to efficiently analyze a large-scale metabolic network without enumerating the whole set of EMs/Expas. The cEM analysis provides a method, including the EM decomposition and maximum entropy principle (MEP), that can efficiently help to computations of complex metabolic systems (Badsha et al,., 2014). The general framework of cEM analysis is shown in figure 3.1.

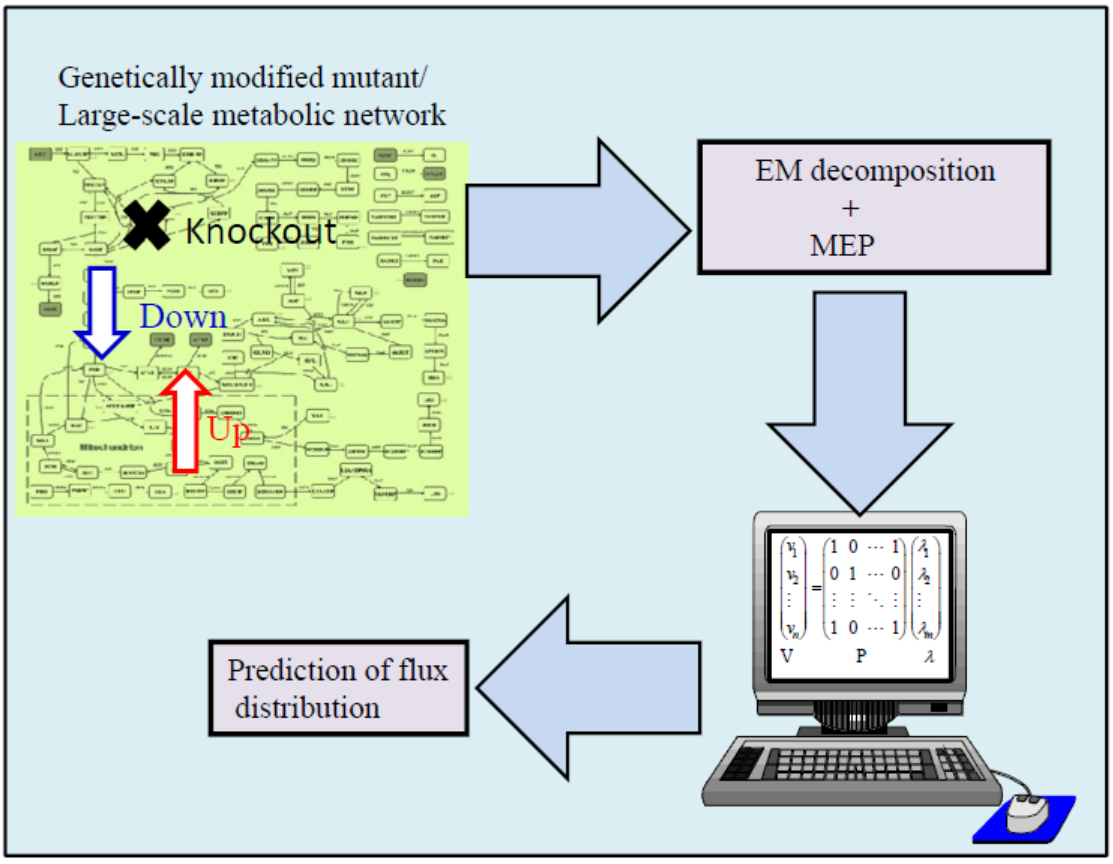


Figure 3.1: The general framework of cEM analysis.

3.2 Complementary Elementary Mode Algorithm

The cEM analysis presents a new method to analyze complex metabolic networks, it combines by three steps: generated many flux distributions by flux balance analysis (FBA) as the input data necessary for the EM decomposition method, reduce the number of EM or decompose the EM (we called cEM) by the EM decomposition method, and flux prediction by maximum entropy principle (MEP). The algorithm for cEM analysis in a given steady-state metabolic network is presented by a flow chart as shown in figure 3.2. The details of cEM method are as follows:

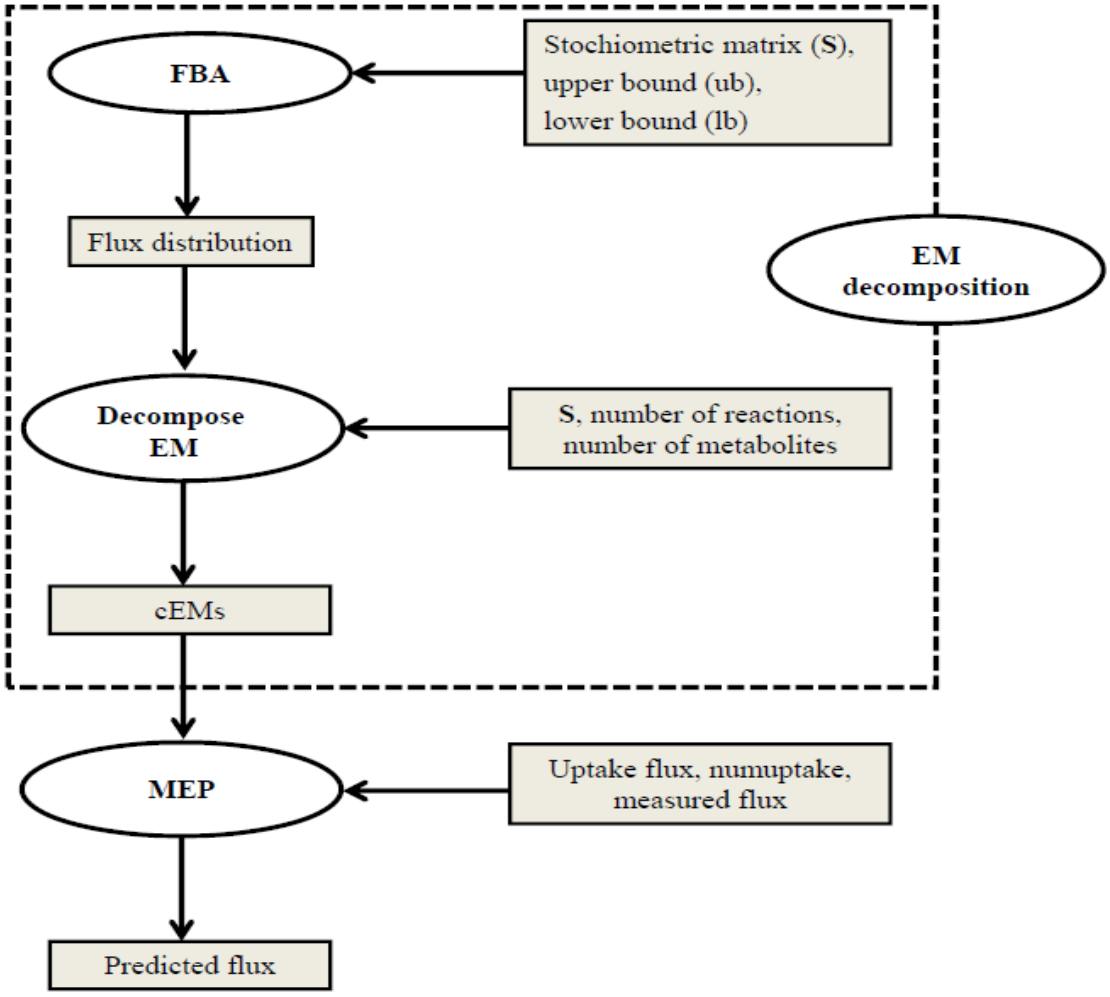


Figure 3.2: A flow chart of cEM analysis. The gray square boxes are the data.The ovals are the algorithms.

3.2.1 Flux Balance Analysis

First, we employ the flux balance analysis (FBA) to determine or generating many metabolic flux distributions that the EM decomposition method necessary as the input file (Figure 3.2). We maximize and minimize each flux distribution at the steady-state level for estimating multiple solutions of metabolic flux distributions, as the following linear programming (LP) problem:

$$
\begin{aligned}
&\text{for} \quad i = 1, 2, 3, \ \ldots, \ n \\
&\text{Maximize} \ \ \mathbf{v}_i = (v_1, v_2, ..., v_n)^t \\
&\text{Subject to: } \mathbf{S} \cdot \mathbf{v} = 0 \qquad \begin{cases} -1000 \le v_i^{rev} \le 1000 \\ 0 \le v_i^{irrev} \le 1000 \end{cases} \\
&\text{for} \quad i = 1, 2, 3, \ \ldots, \ n \\
&\text{Minimize} \ \ \mathbf{v}_i = (v_1, v_2, ..., v_n)^t \\
&\text{Subject to: } \mathbf{S} \cdot \mathbf{v} = 0
\end{aligned}
\tag{3.1}
$$

Where, **v** is the flux vector with a total number of reactions is *n* and **S** is the stoichiometric matrix. The aforementioned optimization, LP problem is performed using MATLAB (Mathworks Inc., Natick, MA). At the most, the employed FBA method can estimate two sets of fluxes such that, 2× the flux number to be estimated.

3.2.2 EM Decomposition of Metabolic Networks

Second, we employ the EM decomposition method (Ip et al., 2011) to decompose the EMs, which determines the major EMs or linear combinations of EMs responsible for predicting the metabolic flux distributions, because computation of the full set of EMs in large-scale metabolic networks is still challenging and sometimes impossible due to its underlying combinatorial complexity. Since EM decomposition are not unique, thus the goal of this method is to assist in the biological interpretation with obtaining a valid

decomposition rather than any specific decomposition. The EM decomposition method does not pick up all EMs, but finds the decomposed pathways that can maintain the metabolic flux at the steady-state level. Those decomposed pathways are named cEMs. The space spanned by the cEMs is included in the space spanned by all the EMs, or cEMs are a subset of EMs.

To obtain cEMs, the EM decomposition method would require many different flux distributions as input. Thus, FBA is used to estimate the entire flux distributions that the EM decomposition method requires as input. The EM decomposition method is an iterative method, first selects the reaction with non-zero flux from a given flux distribution (those are already estimated by FBA) and then applies the mixed integer linear programming (MILP) to find the cEMs that are contained by both the selected and given distributions. Then, it determines the contribution of the cEMs to the given flux distribution. Subtracting the contribution from the given flux distribution presents an updated flux distribution. At the next iteration step, the updated flux distribution is used as the input. This process is repeated until the updated flux distribution approaches to zero. This algorithm takes as input a flux distribution $\boldsymbol{v}$ in the feasible set of optimization problem in equation (3.1) with the FBA model (e.g., $\mathbf{S}$, n, q) (Figure 3.2) and outputs set of cEMs $\{\boldsymbol{p}^{(k)}\}$, which generates $\boldsymbol{v}$ as follows:

$$\mathbf{v} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathbf{p}^{(k)} \boldsymbol{\lambda}_{\mathrm{k}}, \tag{3.2}$$

where $\boldsymbol{\lambda}_k$ is the cEM coefficient, K is the number of cEMs, and q is the number of metabolites. Detail explanation of EM decomposition is described elsewhere (Ip et al., 2011).

3.2.3 Maximum Entropy Principle

In the statement of the problem section, we privilege that many organisms still does not provide any specific biological objective function for estimating the coefficients of EMs/Expas/cEMs and EMs can be described by different scalar products or many possible vectors of each EM, but the predicted flux distributions must be independent of them. The maximum entropy principle (MEP) is convenient in cases where no biological objective function is available and does not depend on scalar products of each EM/Expa/cEM (Zhao and Kurata, 2009a and 2010; Badsha et al., 2013). MEP is derived from Shannon's information theory and is widely used in physics, chemistry, and Bioinformatics for gene expression (Lezon et al., 2006) and sequence analysis (Capra and Singh, 2007; Martin et al., 2005). MEP can widely be implemented in metabolic network analysis for flux predictions (Badsha et al., 2014).

Third, the MEP is used as an objective function to estimate the coefficients of EMs/Expas/cEMs (Figure 3.2). Usually, the metabolic flux distribution at steady-state level can be decomposed onto EMs/Expas/cEMs as follows:

$$\mathbf{v}_d = \mathbf{P}_d \cdot \lambda \tag{3.3}$$

$\mathbf{P}_d$ is the sub-matrix of EM/Expa/cEM matrix in which the determined fluxes are represented by the rows and the EMs/Expas/cEMs by the columns. $\mathbf{v}_d$ is the flux vector with the determined reactions and λ is the coefficients of EMs/Expas/cEMs. Shannon's Entropy is denoted by I and defined as follows:

$$I = -\sum_{j=1}^{ne} \rho_j \log \rho_j \quad , \tag{3.4}$$

where, ρ_j is the probability of EM/Expa/cEM and $\sum_{j=1}^{ne} \rho_j = 1$; *ne* is the the number of EMs/Expas/cEMs. The probability of each EM/Expa/cEM is presented as follows:

$$\rho_j = \frac{1}{v_{\text{substrate uptake}}} \mathbf{P}_{\text{substrate uptak},j} \cdot \lambda_j \left(\sum_{j=1}^{ne} \rho_j = 1 \right) \quad , \tag{3.5}$$

where, $v_{\text{substrate uptake}}$ is the substrate uptake of a flux, $\mathbf{P}_{\text{substrate uptake},j}$ is the element of the *j*th EM/Expa/cEM. Assuming that the contribution of the internal loops ($\mathbf{P}_{\text{substrate uptake},j} = 0$) is neglected based on loop law thermodynamic constraints (price et al., 2006). The internal loop has two reactions, *South* and *frd* in the employed metabolic network model of *E. coli* (See reaction 29 and 30 in the appendix A). The probability of each EM/Expa/cEM, ρ_j is provided by solving the following MEP optimization problem as follows:

$$\text{Maximize} \quad -\sum_{j=1}^{ne} \rho_j \log \rho_j \tag{3.6}$$

$$\text{subject to} \quad \sum_{j=1}^{ne} \rho_j = 1 \tag{3.7}$$

$$\sum_{j=1}^{ne} \rho_j \mathbf{x}_{r,j} = v_r \quad (r = 1, 2, ..., nd) \quad , \tag{3.8}$$

where, v_r is the *r*-th determined flux and *nd* is the number of the determined fluxes. New matrix $\mathbf{x}_{r,j}$, converted from EM/Expa/cEM matrix $\mathbf{P}_d$. There are mainly three steps to apply the MEP objective function for the flux prediction as given below:

Step-1: Normalization of EM/Expa/cEM matrix is given by following form as follows:

$$\mathbf{x}_{r,j} = \begin{cases} \dfrac{v_{substrate\ uptake}}{p_{substrate\ uptake,j}} p_{r,j} & (if\ p_{substrate\ uptake,j} \neq 0) \\ 0 & (if\ p_{substrate\ uptake,j} = 0) \quad , \end{cases} \tag{3.9}$$

where, $p_{r,j}$ is the element of the *r*-th determined flux and *j*-th EM/Expa/cEM.

Step-2: Convert the nonlinear program problem to make a nonlinear equation as follows:

$$\frac{\sum_{j=1}^{ne} x_{r,j} \exp(-\sum_{r=1}^{nd} \psi_r x_{r,j})}{\sum_{j=1}^{ne} \exp(-\sum_{r=1}^{nd} \psi_r x_{r,j})} - v_r = 0 \quad (r = 1, 2, ..., nd) \tag{3.10}$$

The nonlinear equation (3.7) for ψ could be solved by mmfsolve or fsolve in Matlab (Hasbun, 2008).

Step-3: The probabilities of EM/Expa/cEM, ρ_j and coefficient of EM/Expa/cEM, λ_j are calculated by the following two equations as follows:

$$\rho_j = \frac{\exp(-\sum_{r=1}^{nd} \phi_r x_{r,j})}{\sum_{j=1}^{ne} \exp(-\sum_{r=1}^{nd} \phi_r x_{r,j})} \tag{3.11}$$

$$\lambda_j = \begin{cases} \dfrac{v_{substrate\ uptake}}{p_{substrate\ uptake,i}} \rho_j & (if\ p_{substrate\ uptake,j} \neq 0) \\ 0 & (if\ p_{substrate\ uptake,j} = 0) \end{cases} \tag{3.12}$$

Finally, the metabolic flux distribution of the target model is predicted as follows:

$$\mathbf{v}^{\text{target}} = \mathbf{P} \cdot \lambda^{\text{target}} \tag{3.13}$$

Detail explanation of MEP is described elsewhere (Zhao and Kurata, 2009a, 2010).

3.4 Quantitative Contributions

The quantitative contribution to input flux is the measure of how much each cEM/EM is responsible for the input flux, because all of the cEM/EM go through the input flux (Badsha et al., 2014). After applying the MEP step, we calculate the quantitative contribution of each cEM/EM to input flux. The quantitative contribution to input flux is defined as:

$$\text{Quantitative contributions} = \lambda_j \times \mathbf{P}\left(numuptake, j\right), \tag{3.14}$$

where, $\boldsymbol{\lambda}$ is the coefficient vector of cEMs/EMs, j is the interest index of cEMs/EMs, and *numuptake* is the row index that corresponds to the uptake or input flux.

We could be checking the critical cEMs to guarantee the quality of the metabolic flux prediction and to validate the algorithm. The critical number of cEMs is the minimum number that can accurately predict the flux distributions, at which the prediction difference (Equation 3.16) almost converges with respect to the number of cEMs. We rank the total unique cEMs for finding a critical number of cEMs according to their quantitative contributions to input flux.

3.5 Prediction Accuracy

The prediction error in the metabolic flux distributions by EM/Expa/cEM analyses is defined as follows:

$$\text{Predection error} = \sqrt{\frac{1}{nd}\sum_{r=1}^{nd}\left(\mathrm{v}_{r,\text{predection}} - v_{r,\exp}\right)^2}, \tag{3.15}$$

where, $\mathrm{v}_{r,\text{predection}}$ is the predicted flux for the r-th reaction, $v_{r,\exp}$ is the experimental data of the r-th reaction, and nd is the number of the determined fluxes. The prediction difference in the metabolic flux distributions between by the cEM and EM/Expa analyses are defined as follows:

$$\text{Predection difference} = \sqrt{\frac{1}{nu}\sum_{g=1}^{nu}\left(\mathrm{v}_{g,\text{cEM}} - v_{g,\text{EM/Expa}}\right)^2}, \tag{3.16}$$

where, $\mathrm{v}_{g,\text{cEM}}$ and $v_{g,\text{EM/Expa}}$ are the predicted fluxes for the g-th reaction by cEM and EM/Expa analyses, respectively, and nu is the number of the unmeasured fluxes.

3.6 Implementation

The EMs were calculated by the CellNetAnalyzer (CNA), which is a package for MATLAB and provides a comprehensive and user-friendly environment for structural and functional analysis of biochemical networks and used for calculation of EMs and Expas (Klamt et al., 2007) and elementary flux mode tool (efmtool) (Terzer and Stelling, 2007), which is implemented in Java and has been integrated into MATLAB. The Expas were calculated by the CNA with an optional setting (Klamt et al., 2007). The EM decomposition method was implemented using MATLAB with Gurobi 4.0. The Gurobi optimizer is necessary for the EM decomposition method. The optimization of the coefficients of EMs/Expas/cEMs and the prediction of the flux distributions were performed in MATLAB R2014a. The employed computer is the Dell- Windows 7 Professional (Intel-R Core-TM i7-3770; CPU 3.40 GHz; Memory-RAM, 8.00 GB).

3.7 Metabolic Network Models

To investigate the applicability of cEM analysis, compare with EM analysis, we used two medium-scale metabolic networks of *E. coli* (Hua et al., 2006) and one genome-scale metabolic network of head and neck cancer cells (Agren et al., 2012). Details in the metabolic network models are shown in Table 3.1 (The reactions and metabolites of *E. coli* models are shown in Appendix A). The model-I has 140 metabolites and 156 reactions, including reaction 1–5, 7–104, and 107–159 (Figure B.1); model-II has 140 metabolites and 157 reactions, including reaction 1–104 and 107–159 (Figure B.2). The experimental flux distributions for *E. coli* were determined by ^{13}C tracer experiments (Hua et al., 2006). The model-I indicates the *E. coli pta-adhE-pfkA-glk* gene knockout mutant undergoing adaptive evolution for 30 and 60 days under anaerobic condition. The model-II indicates the *E. coli pta-pfkA* gene knockout mutants undergoing adaptive evolution for 30 and 60 days. GMF is

applied to predict the metabolic flux distributions of the genetic mutants of model-I and model-II, where *pgi* and *ppc* genes are over-expressed (The relative enzyme activity ratio of a mutant versus wild type is more than 1) and, *aceA* and *zwf* genes are under-expressed (The relative enzyme activity ratio of a mutant versus wild type is less than 1). Model-III indicates the genome-scale metabolic network of head and neck cancer cells with 4487 metabolites and 2931 reactions. No experimental flux data are available for model-III.

Table 3.1: Details for two metabolic network models of *E. coli* and a genome-scale metabolic network model of head and neck cancer cells (Badsha et al., 2014).

Model	I	II	III
O_2	Anaerobic	Anaerobic	
Substrates	Glucose	Glucose	
Products	Acetate, ethanol, succinate, glycerol, formate, lactate, CO_2	Acetate, ethanol, succinate, glycerol, formate, lactate, CO_2	
# Reactions	156	157	2931
# Metabolites	140	140	4487

3.8 Conclusion

A metabolic flux distribution is the consequence of complex regulations at the enzyme level. A regulatory network is built from the interaction of various levels, such as gene expression, protein-protein interaction and intracellular metabolic reactions. In this chapter, we presented a fast and efficient algorithm, complementary elementary modes (cEMs) analysis to efficiently analyze a large-scale metabolic network. The cEM analysis consists of the FBA, EM decomposition method, and MEP. To rationally design of metabolic networks, the cEMs are very effective in integration of gene expression data into large-scale models by GMF.

Chapter 4

Results and Discussions

4.1 Introduction

We consider both artificial and real metabolic network analyses for comparison between the performance and applicability of the cEM and existing EM/Expa algorithms.

4.2 Simulation Study

To investigate the performance of the cEM analysis in a comparison of the existing EM/Expa analysis, we consider a synthetic metabolic network as shown in figure 4.1.

4.2.1 Artificial Metabolic Network

To explain the procedure of the cEM analysis, we applied to the following plain artificial metabolic network model (figure 4.1). Where X1, X2 and X3 (cycles) are the internal metabolites that need to fulfill a steady-state, while $X1_{ex}$, $X2_{ex}$ and $X3_{ex}$ (squares) are the external metabolites that need not be balanced in this scheme. The input flux of v_1 is fixed to 100. We have to predict the unknown fluxes $(v_2, v_3, v_4, v_5, v_6, v_7)$.

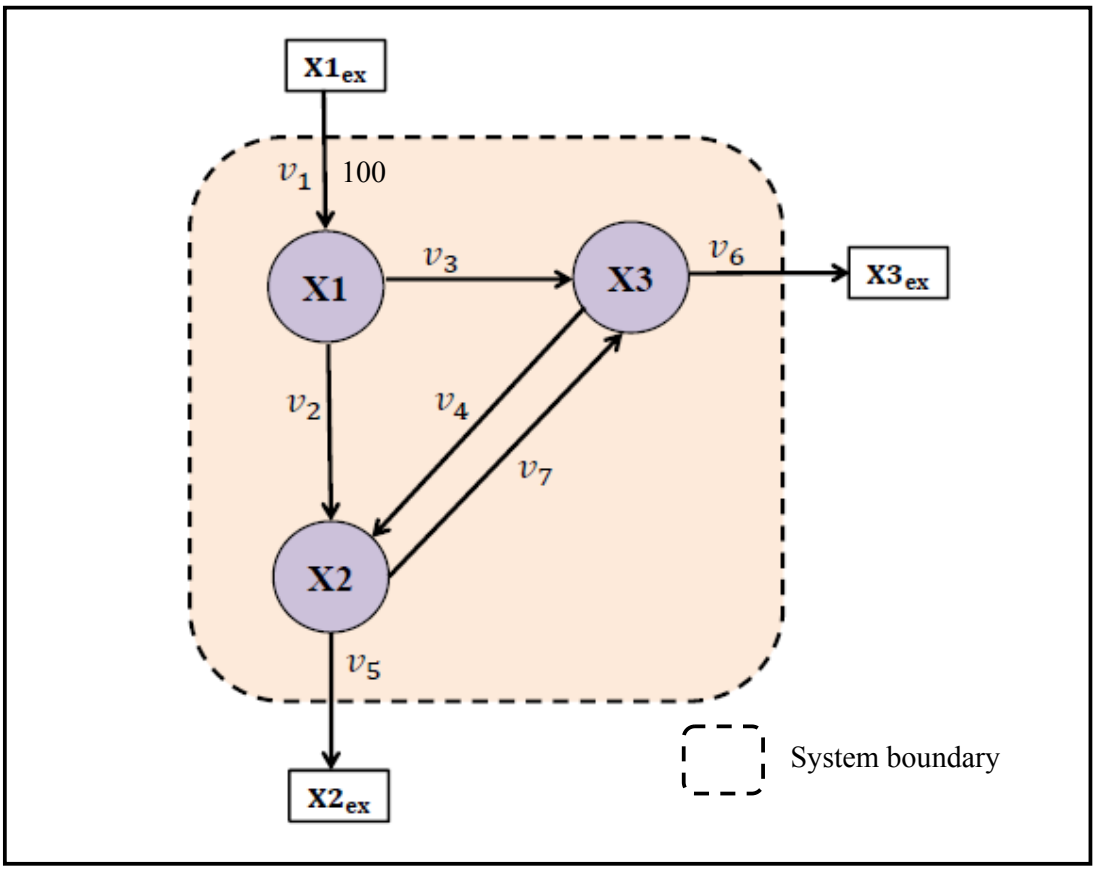


Figure 4.1: Synthetic metabolic network

The above synthetic metabolic model can be represented by a stoichiometric matrix, **S**. For the given model, the metabolites are represented by the rows of **S** and the columns of **S** correspond to the reactions in a network. The coefficient of stoichiometric matrix **S**, we use positive (+) sign for a metabolite is formed (produced) by the reaction and negative (-) sign for a metabolite is consumed by the reaction. Then, we can express **S** for the synthetic metabolic network as follows:

$$\mathbf{S} = \begin{array}{c} \\ \text{X1} \\ \text{X2} \\ \text{X3} \end{array} \begin{array}{c} \begin{array}{ccccccc} \text{v}_1 & \text{v}_2 & \text{v}_3 & \text{v}_4 & \text{v}_5 & \text{v}_6 & \text{v}_7 \end{array} \\ \begin{pmatrix} 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 & -1 & 1 \end{pmatrix} \end{array} \tag{4.1}$$

At the steady-state level, the equations are given by:

$$\mathbf{S} \cdot v = 0 \tag{4.2}$$

EM/Expa analysis

We get 5 EM/Expa by applying the EM/Expa analysis of the synthetic metabolic network. We found that the same number of EM and Expa, because there is no reversible reaction are involved in the synthetic metabolic network model that shown in figure 4.1. The EM/Expa analysis results are illustrated in the figure 4.2.

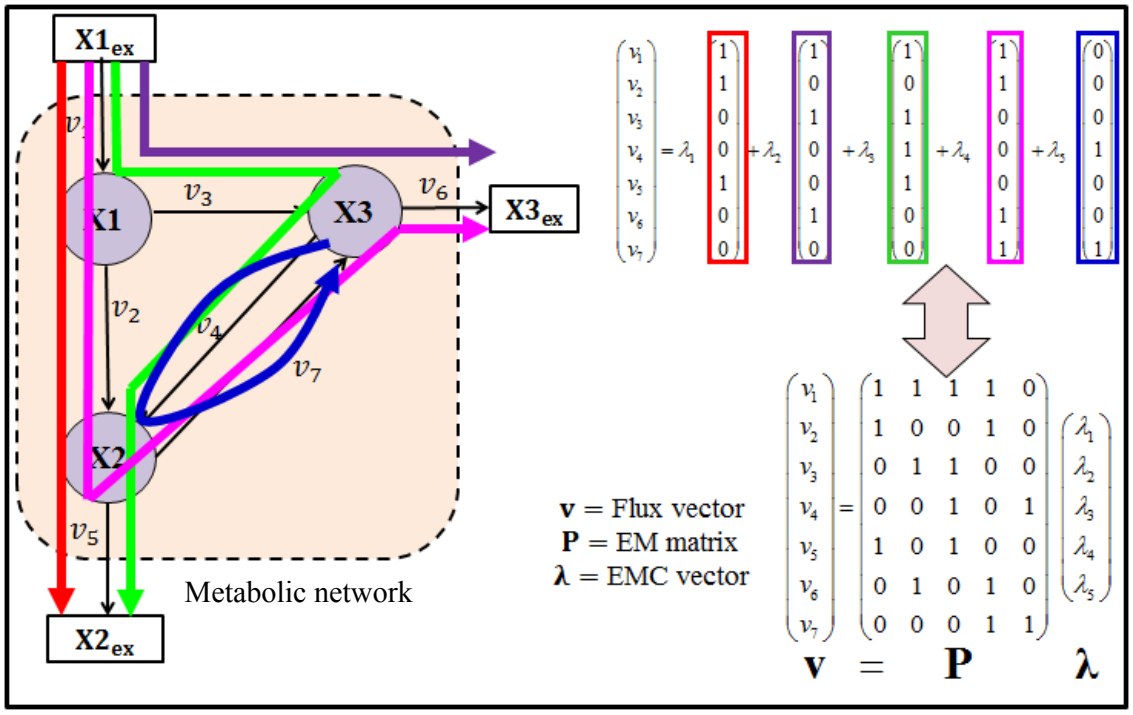


Figure 4.2: EM/Expa analysis of synthetic metabolic network.

Scalar product of EM analysis

The EM can be described by many possible vector or scalar products of each EM. To describe the scalar product problem of EM analysis, we consider the synthetic metabolic network as shown in figure 4.1. The following equations 4.3, 4.4 and 4.5 are the possible vector or scalar products of each EM.

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \end{pmatrix} = \lambda_1 \begin{pmatrix} 2 \\ 2 \\ 0 \\ 0 \\ 2 \\ 0 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 3 \\ 0 \\ 3 \\ 0 \\ 0 \\ 3 \\ 0 \end{pmatrix} + \lambda_3 \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \lambda_4 \begin{pmatrix} 4 \\ 4 \\ 0 \\ 0 \\ 0 \\ 4 \\ 4 \end{pmatrix} + \lambda_5 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 5 \\ 0 \\ 0 \\ 5 \end{pmatrix} = \begin{pmatrix} 2 & 3 & 1 & 4 & 0 \\ 2 & 0 & 0 & 4 & 0 \\ 0 & 3 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 5 \\ 2 & 0 & 1 & 0 & 0 \\ 0 & 3 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 & 5 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \end{pmatrix} \quad (4.3)$$

We can write the above equation 4.3 as the following from as follows:

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \end{pmatrix} = 2\lambda_1 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + 3\lambda_2 \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} + \lambda_3 \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + 4\lambda_4 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} + 5\lambda_5 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2\lambda_1 \\ 3\lambda_2 \\ \lambda_3 \\ 4\lambda_4 \\ 5\lambda_5 \end{pmatrix} \quad (4.4)$$

Let, $\lambda_1' = 2\lambda_1, \lambda_2' = 3\lambda_2, \lambda_3' = \lambda_3, \lambda_4' = 4\lambda_4$ and $\lambda_5' = 5\lambda_5$, then we can write as follows:

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \end{pmatrix} = \lambda_1' \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \lambda_2' \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} + \lambda_3' \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \lambda_4' \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} + \lambda_5' \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1' \\ \lambda_2' \\ \lambda_3' \\ \lambda_4' \\ \lambda_5' \end{pmatrix} \quad (4.5)$$

Compare with existing EM/Expa analysis, we apply the cEM analysis, which consists of three steps as follows (details are described in the materials and method section):

<u>Step-1</u>: Flux balance analysis (FBA)

<u>First</u>, we employ the FBA method to determine the possible metabolic flux distributions, which are necessary for the EM decomposition method. In the FBA, multiple solutions of

metabolic flux distributions are calculated by maximizing and minimizing each flux in a given steady-state as follows:

$$
\begin{aligned}
&\text{Maximize } \mathbf{v}_i = (v_2, v_3, v_4, v_5, v_6, v_7)^t \\
&\text{for } \quad \text{i} = 2, 3, 4, 5, 6, 7 \quad \text{and } v_1 = 100 \\
&\text{Subject to: } \mathbf{S} \cdot \mathbf{v} = 0 \quad \{1e\text{-}8 \le \mathbf{v}_i \le 1000 \quad \text{(all reactions are irreversible)} \qquad (4.6) \\
&\text{Minimize } \mathbf{v}_i = (v_2, v_3, v_4, v_5, v_6, v_7)^t \\
&\text{for } \quad \text{i} = 2, 3, 4, 5, 6, 7 \quad \text{and } v_1 = 100 \\
&\text{Subject to: } \mathbf{S} \cdot \mathbf{v} = 0
\end{aligned}
$$

The upper bound (ub) for all reactions is set to 1000 and the lower bound (lb) is set to 1e-8 (Since all reactions are irreversible). Input flux v_1 is fixed to 100 $(\text{for } \text{v}_1, \text{ub}_1 = 100 \text{ and } \text{lb}_1 = 100)$ and the other 12 sets of metabolic flux distributions are estimated by maximizing and minimizing each flux (solve equation (4.6)) as the following table 4.1.

Table 4.1: 12 sets of flux distributions are estimated by solving equation 4.6.

V_2		V_3		V_4		V_5		V_6		V_7	
Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min
100	100	100	100	100	100	100	100	100	100	100	100
100	0	0	100	0	50	66.67	33.33	33.33	66.67	100	50
0	100	100	0	100	50	33.33	66.67	66.67	33.33	0	50
0	33.33	33.33	0	1000	0	33.33	0	0	33.33	900	0
66.67	33.33	33.33	66.67	100	50	100	0	0	100	0	50
33.33	66.67	66.67	33.33	0	50	0	100	100	0	100	50
33.33	0	0	33.33	900	0	0	33.33	33.33	0	1000	0

Then, we have to select the unique or independent flux. Because, if we use the same flux for the input file of EM decomposition then we get the same cEM. Therefore, we take seven independent or unique sets from the 12 sets, as shown by the following table as follows:

Table 4.2: Seven independent / unique sets from table 4.1.

1	2	3	4	5	6	7
100	100	100	100	100	100	100
100	0	0	50	66.67	33.33	100
0	100	100	50	33.33	66.67	0
0	33.33	1000	0	33.33	0	900
66.67	33.33	100	50	100	0	0
33.33	66.67	0	50	0	100	100
33.33	0	900	0	0	33.33	1000

Step-2: Elementary mode (EM) decomposition

Second, we apply the EM decomposition technique, which generates the major EMs or linear combination of EMs (cEMs), which are the responsible for the flux distributions. The EM decomposition method uses mixed integer linear programming (MILP) to find the cEMs that satisfy the steady-state condition for 7 unique sets of flux distributions. We obtain 5 independent or unique cEMs from 7 sets of flux distribution as follows:

$$\text{cEM}(\mathbf{P}) = \begin{pmatrix} \text{cEM1} & \text{cEM2} & \text{cEM3} & \text{cEM4} & \text{cEM5} \\ 0.5774 & 0.5774 & 0.5 & 0.5 & 0 \\ 0.5774 & 0 & 0 & 0.5 & 0 \\ 0 & 0.5774 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0.7071 \\ 0.5774 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5774 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 & 0.7071 \end{pmatrix}_{7\times 5} \tag{4.7}$$

Step-3: Maximum entropy principle (MEP)

***Third*,** we use the MEP as an objective function to estimate the 5 coefficients of cEMs, to avoiding the scalar product problem of EM. By removing a closed pathway, the following

4 cEMs were found enough to estimate the flux distributions, at which the predicted difference is zero. The 4 cEM by the cEM analysis as follows:

$$\text{cEM}=\begin{bmatrix} \text{cEM1} & \text{cEM2} & \text{cEM3} & \text{cEM4} \\ 0.5774 & 0.5774 & 0.5 & 0.5 \\ 0.5774 & 0 & 0 & 0.5 \\ 0 & 0.5774 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0.5774 & 0 & 0.5 & 0 \\ 0 & 0.5774 & 0 & 0.5 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}_{7\times4} \tag{4.8}$$

The metabolic flux distributions predicted by the cEM analysis are given as follows:

Flux distributions
100
50
50
25
50
50
25

4.3 Real Metabolic Network

4.3.1 Flux Predictions

To demonstrate the feasibility of cEM analysis compared with ordinary EM analysis, we calculated the flux distribution of two metabolic networks of *E. coli* (Badsha et al., 2014). In model-I, the cEM analysis generated 202 unique cEMs, while the EM analysis generated 122,126 EMs by using the CNA and efmtool. In model-II, cEM analysis generated 295 unique cEMs and EM analysis did 321,416 EMs.

4.3.1.1 Model-I

In model-I, the flux distributions were estimated by both cEM and EM analyses, as shown in figure 4.3. The predicted flux distributions were compared with 26 experimental fluxes (Hua et al., 2006) as shown in figure 4.3(A, B). Figure 4.3(A) and 4.3(B) shows the flux prediction for 30-day-cultured cells and 60-day-cultured cells, respectively. The 130 unmeasured fluxes predicted by the cEMs were compared with those by the EMs generated by CNA in figure 4.3(C, D). Figure 4.3(C) and 4.3(D) shows the 130 unmeasured predicted flux for 30-day-cultured cells and 60-day-cultured cells, respectively. The prediction errors as defined by equation (3.15), are calculated by cEM and EM analyses for 30 and 60-day-cultured cells as shown in figure 4.3(E). From these figures 4.3(A-B), we have seen that, the predicted flux distribution by the cEMs was very consistent with the experimental data and also consistent with the predicted flux distribution by the EMs. Moreover, figure 4.3(B-C) to confirm that the unmeasured predicted fluxes was very closely related to the cEM and EM analyses. Here, 29 critical cEMs were selected out of 202, as shown in figure 4.3(F). To find a critical number of cEMs, we ranked the total 202 cEMs according to their quantitative contributions to input flux and estimated the prediction differences by equation (3.16). The 29 critical cEMs, which are only a 2.4×10^{-4} portion of the total EMs, were found enough to estimate the flux distributions, at which the prediction difference almost converged as shown in figure 4.3(F).

4.3.1.2 Model-II

In model-II, the flux distributions were estimated by both cEM and EM analyses by CNA, as shown in figure 4.4. In figure 4.4(A, B) the predicted flux distributions were compared with 26 experimental fluxes (Hua et al., 2006). The predicted fluxes for 30-day-cultured cells and 60-day-cultured cells are shown in figures 4.4(A) and 4.4(B), respectively. In figure 4.4(C, D), the 131 unmeasured fluxes predicted by the cEMs were compared with

those by the EMs. The predicted 131 unmeasured fluxes for 30-day-cultured cells and 60-day-cultured cells, are shown in figure 4.4(C) and 4.4(D), respectively. The prediction errors for cEM and EM analyses are shown in figure 4.4(E). In analysis of figures 4.4(A-B), we have seen that the predicted flux distributions by the cEMs were consistent with the experimental data and also consistent with the predicted flux distribution by the EMs. The predicted unmeasured fluxes also very consistent by the cEM and EM analyses. In the same manner as figure 4.3(F), we ranked the total 295 cEMs due to their quantitative contributions to input flux and calculated the prediction difference as shown in figure 4.4(F). The 35 critical cEMs, which are only a 1.1×10^{-4} portion of the total EMs, were found enough to estimate the flux distributions, at which the prediction difference almost converged as shown in figure 4.4(F).

In both model-I and model-II, the prediction errors by the cEMs were comparable to those by the EMs, presenting that cEMs are effective in prediction of the flux distribution and that a very small portion of the total EMs are enough to predict flux distributions, at which the prediction difference almost converged. On the other hand, the prediction difference between cEM and EM analyses did not become zero, as shown in figure 4.3(F) and figure 4.4(F), suggesting that many EMs with a small value of EMCs can a little affect the flux distributions. The difference would be also caused by the fact that a set of the critical cEMs are different from that of the EMs, as shown in Section 4.3.3.

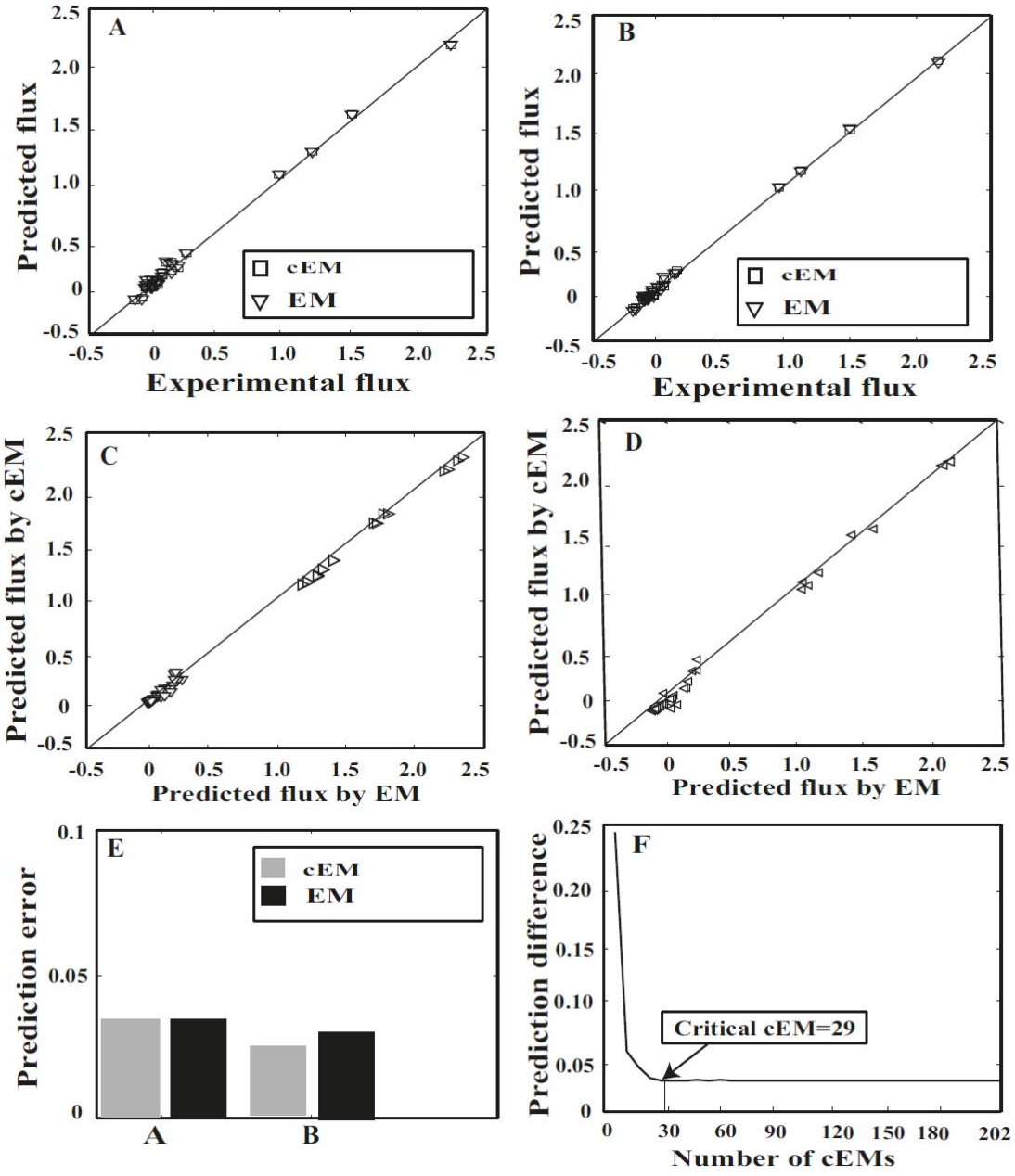


Figure 4.3: Flux distributions predicted by cEM and EM analyses for model-I of *E. coli* mutants. (A and B) The predicted flux distributions are compared with 26 experimental fluxes for 30-day-cultured cells (A) and 60-day-cultured cells (B). (C and D) The predicted flux distribution by the cEMs is compared with that by the EMs for 30-day-cultured cells (C) and 60-day-cultured cells (D), where 130 unmeasured fluxes are estimated. (E) The prediction errors are calculated by cEM (gray) and EM (black) analyses for 30 and 60-day-

cultured cells (A and B). (F) The prediction difference is plotted with respect of the number of cEMs.

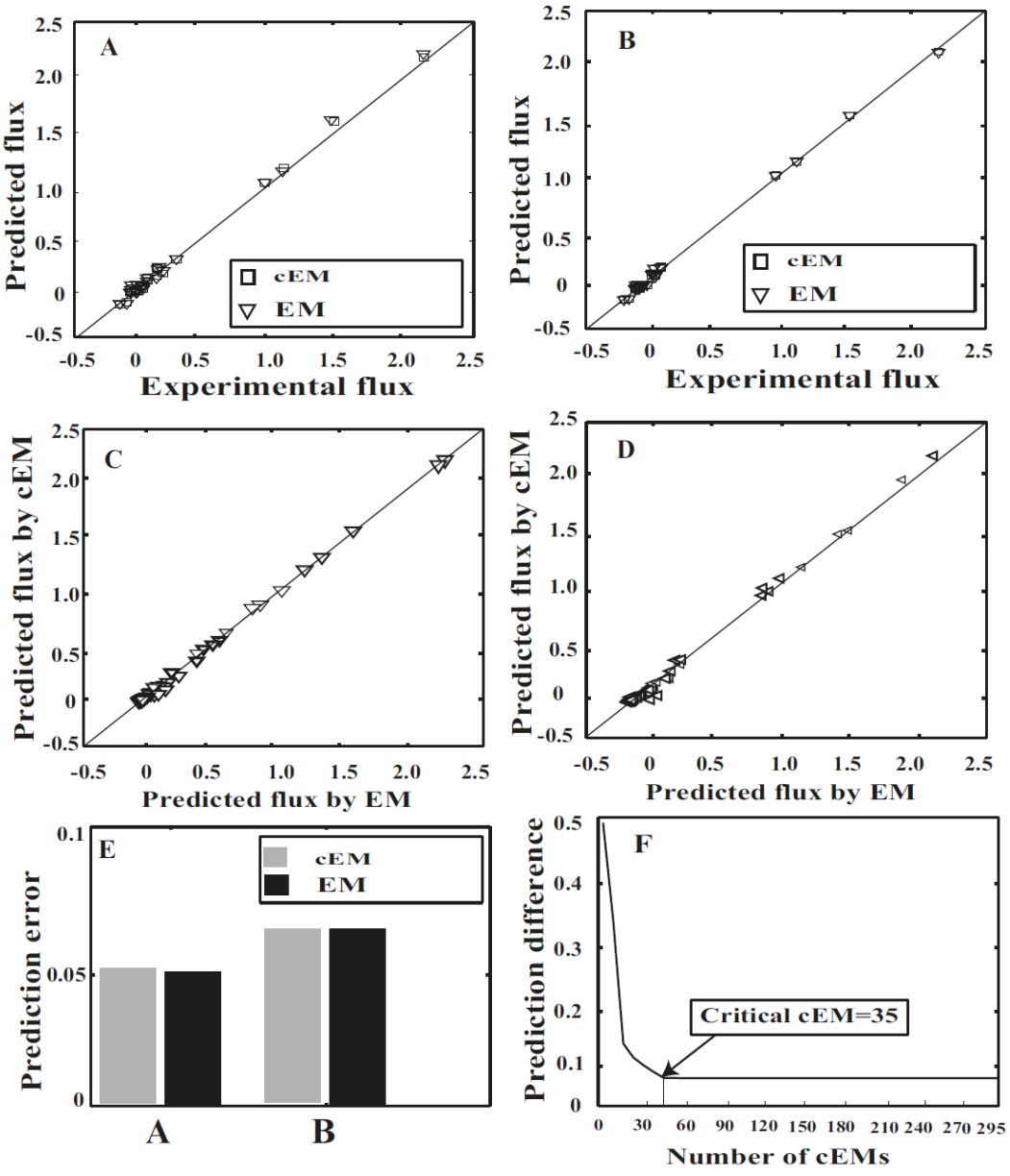


Figure 4.4: Flux distributions predicted by cEM and EM analyses for model-II of *E. coli* mutants. (A and B) The predicted flux distributions are compared with 26 experimental fluxes for 30-days-cultured cells (A) and 60-day-cultured cells (B). (C and D) The predicted flux distribution by the cEMs is compared with that by the EMs for 30-day-cultured cells (C)

and 60-day-cultured cells (D), where 131 unmeasured fluxes are estimated. (E) The prediction errors are calculated by cEM (gray) and EM (black) analyses for 30-and 60-day-cultured cells (A and B). (F) The prediction difference is plotted with respect of the number of cEMs.

4.3.2 Statistical Analysis of Prediction Accuracy

A measure of prediction accuracy by statistical analysis is an important process for comparing the two systems. To show a correlation between the predicted and experimental fluxes for model-I and model-II, we performed a statistical analysis using a linear regression model (Badsha et al., 2014). Table 4.3 shows the Pearson's correlation coefficient (r) (Pearson, 1895), the coefficients of determination (R^2) (Steel, and Torrie, 1960) between the experimental and predicted fluxes, and *P* values (Goodman, 1999) for cEM and EM analyses. The *P* values are used for testing the hypothesis that the predicted and experimental flux distributions are uncorrelated. The correlation coefficients range between 0.9634 and 0.9989, the coefficients of determinations range from 0.9281 to 0.9978, and the *P* values from 1.9×10^{-33} to 3.4×10^{-15}. These statistical analyses demonstrate that the correlation coefficient and coefficients of determination are remarkably high and the *P* values are significantly small (i.e., reject the hypothesis at the 5% level of significance), presenting a statistical high consistency between the experimental and predicted fluxes by the cEM and EM analyses.

Table 4.3: The Pearson's correlation coefficient (r), the coefficients of determination (R^2) between the experimental and predicted fluxes and *P* values by the cEM and EM analyses.

Model	Condition	Method	Pearson's Correlation (r)	Coefficients of determination (R^2)	*P* value
Model-I	Wild type	EM	0.9980	0.9960	2.4×10^{-30}
		cEM	0.9982	0.9964	7.4×10^{-31}
	Mutant type	EM	0.9973	0.9947	7.9×10^{-29}
		cEM	0.9975	0.9950	3.5×10^{-30}
Model-II	Wild type	EM	0.9639	0.9291	2.6×10^{-15}
		cEM	0.9634	0.9281	3.4×10^{-15}
	Mutant type	EM	0.9989	0.9978	1.9×10^{-33}
		cEM	0.9789	0.9878	1.7×10^{-20}

4.3.3 Quantitative Contributions of cEMs and EMs

To make clear the difference between cEM and EM analyses, we characterized the critical 29 and 35 cEMs and their quantitative contributions to the input flux (Badsha et al., 2014). We estimated the quantitative contribution by calculating how much each cEM/EM is responsible for the input flux (Equation 3.14), because all the cEMs and EMs go through the input flux. The employed cEMs and EMs and their quantitative contributions are summarized using Venn-diagrams for model-I and model-II, as shown in figure 4.5.

4.3.3.1 Model-I

To calculate the quantitative contribution to input flux for model-I, the input flux (glucose uptake) was set to 1. The number of cEMs was set to 29 as the same cEMs as used for flux prediction, while that of EMs was set to 122,126. We sorted the cEMs/EMs in the descending order of their contributions, respectively. Top four cEMs and EMs were the same, while their quantitative contributions to input flux were a little bit different between both the cEMs and EMs analyses, as shown in figure 4.6(A). The pathway length and quantitative contributions to input flux with respect to the 4 consistent cEMs/EMs are listed in table 4.4.

The maximum contribution of the cEM analysis with a pathway length of 14 was approximately 0.1680, whereas that of the EM analysis with a pathway length of 15 was 0.2002.

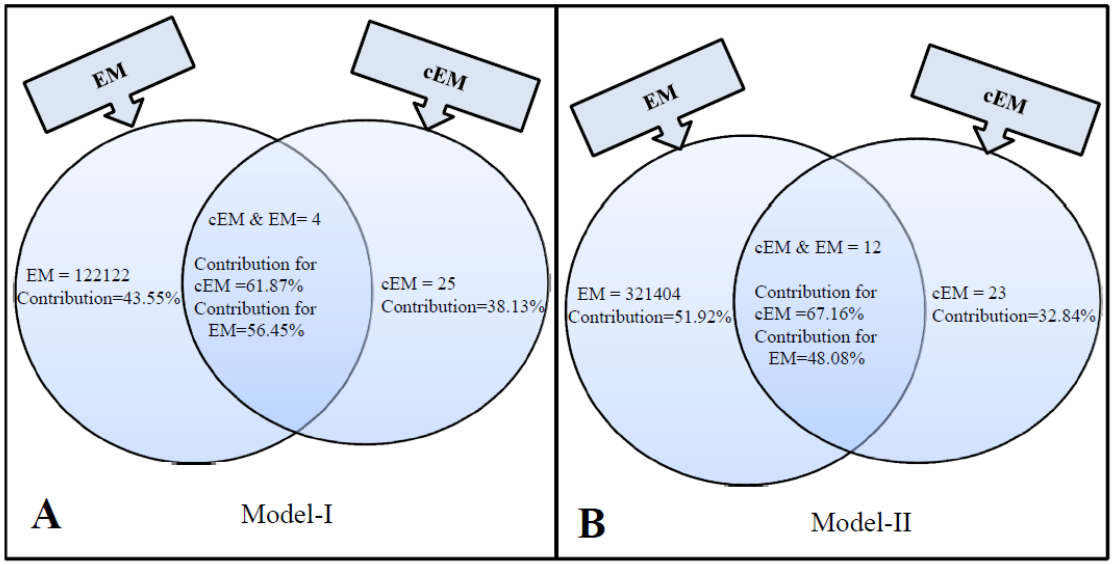


Figure 4.5: Venn-diagram (A) Model-I and (B) Model-II. The employed cEMs and EMs and their quantitative contributions to input flux are illustrated.

The maximum contribution of the cEM analysis with a pathway length of 14 was approximately 0.1680, whereas that of the EM analysis with a pathway length of 15 was 0.2002. In general, the pathway length of biomass formation is long, because it includes many reactions of amino acids, DNA, and RNA syntheses. The pathway length of lactate production is short, as they belong to glycolysis and pyruvate metabolism. The metabolic pathways of the 4 consistent cEMs/EMs are shown in figure B.2 (Appendix B). The 4 cEMs/EMs were not related to biomass formation, but were related to the formation of lactate; one of them was coupled with ATP drain. Thus, the pathway lengths for the consistent cEMs/EMs were short. The cEM with the maximum contribution, which is related

to ATP drain, showed 16.80%; its corresponding EM analysis did 12.14%. These 4 consistent cEMs showed total 61.87% contribution to input flux; the corresponding EMs did 56.45%.

Table 4.4: The four consistent EMs or cEMs. Pathway length and quantitative contributions to input flux by the cEM and EM analyses for model-I. The metabolic pathway maps of the four EMs or cEMs are shown in figure B.2.

Method	Number	Pathway length	Quantitative contributions
EM	2, 4, 5, 3	14,14,15,14	0.1214, 0.1214, 0.2002, 0.1214
cEM	42, 43, 44, 49	14,14,15,14	0.1667, 0.1667, 0.1174, 0.1680

4.3.3.2 Model-II

In model II, the 35 critical cEMs were used, while the number of EMs was 321,416. We sorted the cEMs and EMs in the descending order of their contribution, respectively. As shown in figure 4.6(B), the top twelve cEM and EMs were the same. The 12 consistent cEMs and EMs and their quantitative contributions to the input flux are listed in table 4.5. The maximum contribution of the cEM with a pathway length of 14 is 0.1070; that of the EM is 0.0804. The metabolic pathways of the 12 consistent cEMs / EMs are shown in figure B.3 (Appendix B). The 12 cEMs / EMs are not related to biomass formation, which are related to the formation of ethanol and lactate; some of them are coupled with ATP drain. The 12 consistent cEMs showed total 67.16% contribution, whereas the 12 EMs did 48.08%.

Table 4.5: The 12 consistent EMs or cEMs. Pathway length and quantitative contributions by the cEM and EM analyses for model-II. The metabolic pathway maps of the 12 EMs or cEMs are shown in figure B.3.

Method	Number	Pathway length	Quantitative contributions
EM	3, 4, 6, 7, 8, 11, 12, 14, 26940, 26943, 26945, 26946	14, 14, 14, 14, 14, 17, 17, 17, 15, 16, 16, 16	0.0804, 0.0804, 0.0804, 0.0804, 0.0804, 0.0111, 0.0111, 0.0111, 0.0114, 0.0114, 0.0114, 0.0114
cEM	59, 57, 108, 102, 101, 164, 162, 166, 121, 77, 117, 118	14, 14, 14, 14, 14, 17, 17, 17, 15, 16, 16, 16	0.1070, 0.1062, 0.1069, 0.1062, 0.1062, 0.0288, 0.0288, 0.0286, 0.0132, 0.0132, 0.0132, 0.0132

The quantitative contribution of non-consistent cEMs was relatively large in both the models, while they did not deteriorate the prediction accuracy of the flux distributions (Figures 4.3A-4.3E, 4.4A-4.4E). The prediction difference between cEM and EM analyses (Figures 4.3F, 4.4F) would be caused by the fact that a set of the critical cEMs are different from that of the EMs.

4.3.4 GMF-Predicted Flux Distribution

To further demonstrate the feasibility of cEMs, we applied the cEMs and EM to GMF-predicted flux distributions of the genetically modified model-I and model-II (Badsha et al., 2014).

4.3.4.1 Model-I

In the genetically modified model-I, the flux distribution were estimated with both cEM analysis and EM analysis by CNA, as shown in figure 4.7. In figure 4.7(A, B) the flux distributions predicted by cEM and EM analyses were compared with 26 experimental fluxes (Hua et al., 2006). The GMF-predicted flux distributions by cEM and EM analyses for 30-day-cultured cells and 60-day-cultured cells, are shown in figure 4.7(A) and 4.7(B), respectively. In figure 4.7(C, D) the 130 unmeasured flux distributions predicted by the 29 critical cEMs were compared with those by the EMs. The unmeasured flux distributions for

30-day-cultured cells and 60-day-cultured cells, are shown in figure 4.7(C) and 4.7(D), respectively. The prediction errors as defined in equation (3.7) between the GMF-estimated fluxes by cEM and EM analyses are shown in figure 4.7(E). In figure 4.7F, to find the critical number of cEMs, we sorted the total 202 unique cEMs in the descending order of their quantitative contributions to input flux and calculated the prediction differences between by cEM and EM analyses. The 29 critical cEMs were found enough to estimate the flux distributions for the genetically modified model-I, which were the same EMs as determined by figure 4.3F.

4.3.4.2 Model-II

In the genetically modified model-II, the flux distributions by both cEM and EM analyses are shown in figure 4.8. The 26 predicted fluxes were compared with the experimental flux distributions (Hua et al., 2006), as shown in figure 4.8(A, B). The GMF-predicted flux distributions by cEM and EM analyses for 30-day-cultured cells and 60-day-cultured cells, are shown in figure 4.8(A) and 4.8(B), respectively. In figure 4.8(C, D), the 131 unmeasured flux distribution predicted by the cEMs were compared with those by the EMs. The 131 unmeasured flux distributions for 30-day-cultured cells and 60-day-cultured cells, are shown in figure 4.8(C) and 4.8(D), respectively. Figure 4.8(E) shows the prediction errors by cEM and EM analyses. In figure 4.8F, to find the critical number of cEMs, we sorted the total 295 unique cEMs in the descending order of their quantitative contributions to input flux and calculated the prediction differences between by cEM and EM analyses. The 35 critical cEMs were found enough to estimate the flux distributions for the genetically modified model-I, which were the same EMs as determined by figure 4.4F.

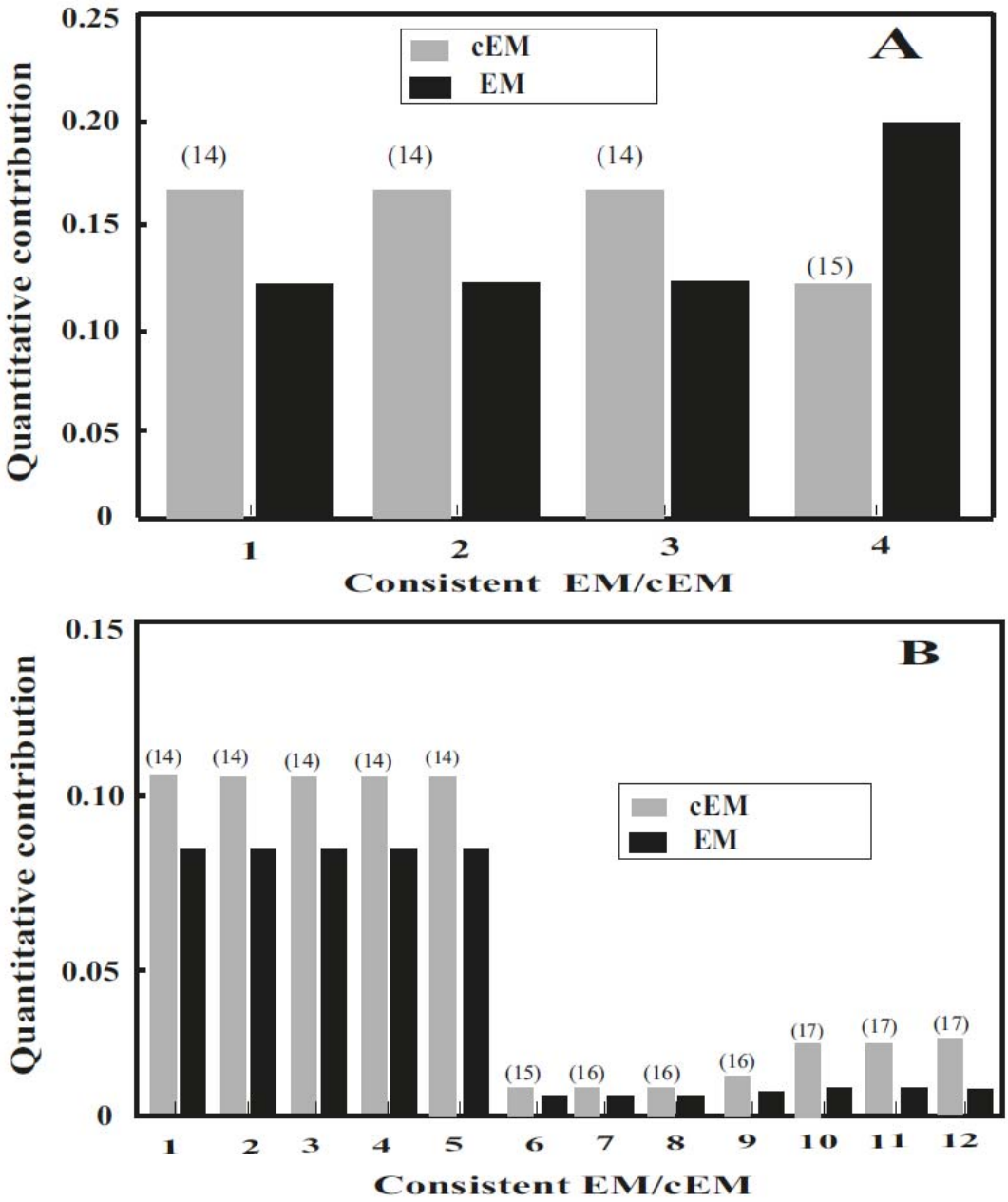


Figure 4.6: The consistent EMs/cEMs and their quantitative contributions to input flux. (A) Model-I and (B) Model-II. The value top of the bar diagram within the bracket indicates the pathway length.

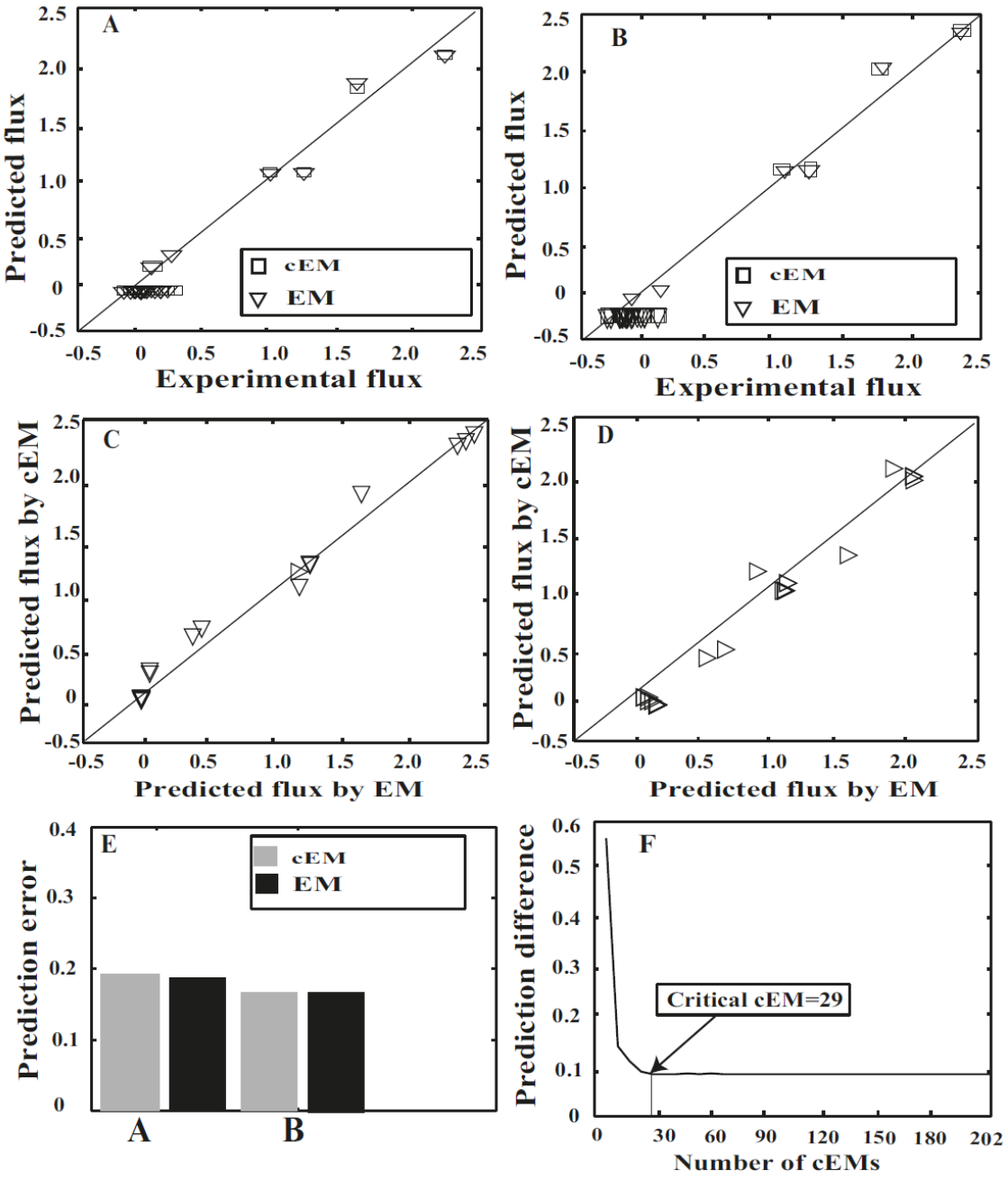


Figure 4.7: Flux distributions predicted by GMF with the cEMs/EMs for the genetically modified model-I. (A and B) The GMF-predicted flux distributions are compared with 26 experimental fluxes for 30-days-cultured cells (A) and 60-day-cultured cells (B). (C and D) The GMF-predicted flux distribution by the cEMs is compared with that by the EMs for 30-day-cultured cells (C) and 60-day-cultured cells (D), where 130 unmeasured fluxes are

estimated. (E) The prediction errors are calculated by cEM (gray) and EM (black) analyses for 30- and 60-day-cultured cells (A and B). (F) The prediction difference is plotted with respect of the number of cEMs.

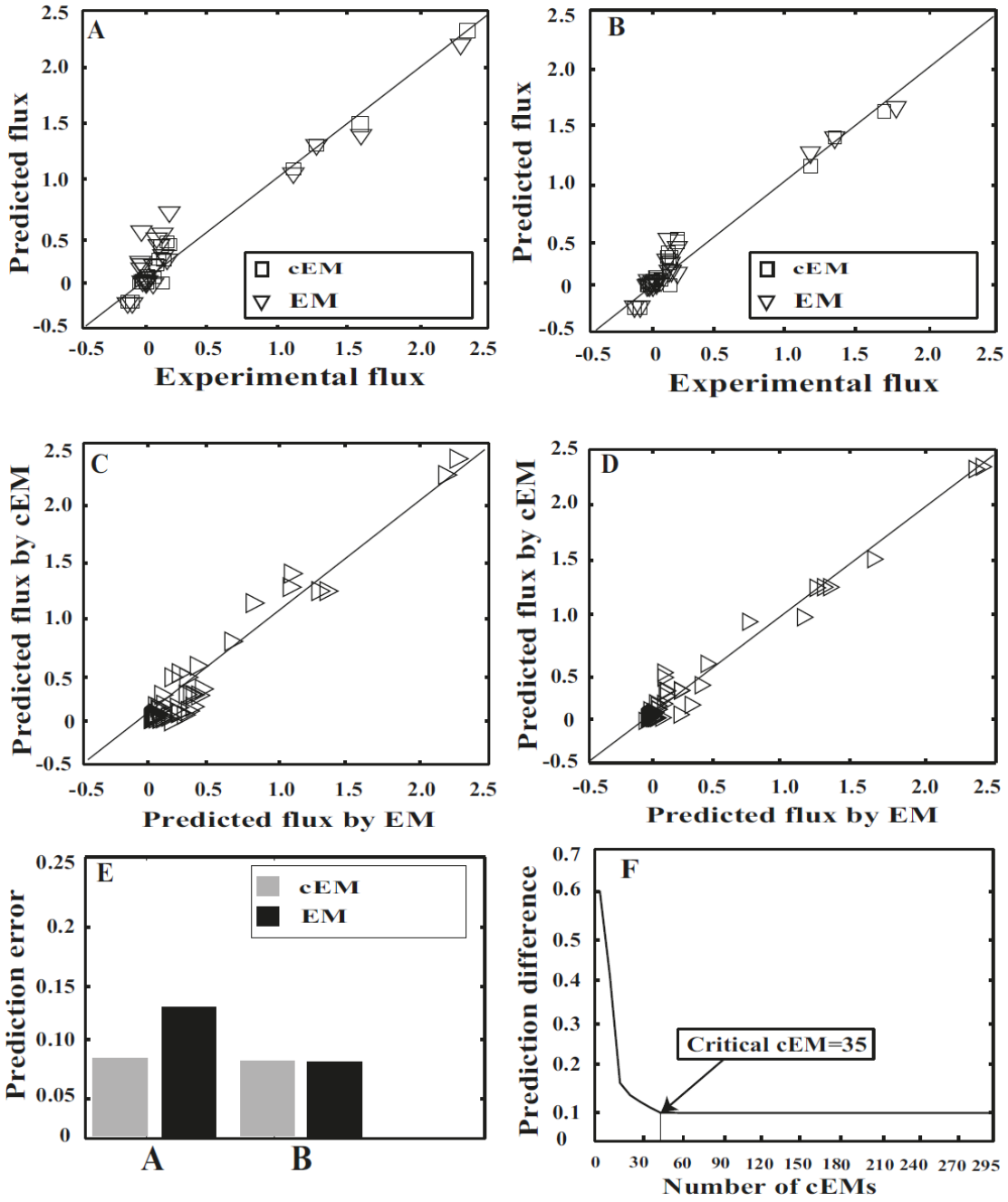


Figure 4.8: Flux distributions predicted by GMF with the cEMs/EMs for the genetically modified model-II. (A and B) The GMF-predicted flux distributions are compared with 26

experimental fluxes for 30-day-cultured cells (A) and 60-day-cultured cells (B). (C and D) The GMF-predicted flux distribution by the cEMs is compared with that by the EMs for 30-day-cultured cells (C) and 60-day-cultured cells (D), where 131 unmeasured fluxes are estimated. (E) The prediction errors are calculated by cEM (gray) and EM (black) analyses for 30- and 60-day-cultured cells (A and B). (F) The prediction difference is plotted with respect of the number of cEMs.

4.3.5 Statistical Analysis of GMF-Prediction Accuracy

The GMF-predicted flux distributions by cEM analysis are rather consistent with those by EM analysis. Some differences may be caused by the fact that the quantitative contributions of the critical cEMs are not the same as those of the EMs (Figures 4.6, 4.7 and 4.8). To statistically characterize their consistency, the Pearson's correlation coefficient (r), the coefficients of determination (R^2) between the experimental and GMF-predicted flux distributions, and the *P* values by cEM and EM analyses are shown in table 4.6. The correlation coefficients range between 0.8723 and 0.9817, the coefficients of determinations range from 0.7610 to 0.9637, and the *P* values from 2.3×10^{-22} to 6.3×10^{-11}. The correlation coefficient and coefficients of determination were remarkably high and the *P* values were significantly small (reject the hypothesis at the 5% level of significance). These results provided statistically significant correlation between the experimental fluxes and the GMF-predicted ones by cEM and EM analyses.

Table 4.6: The Pearson's correlation coefficient (r), the coefficients of determination (R^2) between the experimental and GMF-predicted fluxes and *P* values by the cEM and EM analyses.

Model	Adaptive evolution	Method	Pearson's Correlation (r)	Coefficients of determination (R^2)	*P* value
Model-I	30 days	EM	0.9496	0.9017	3.5×10^{-16}
		cEM	0.9380	0.8798	6.5×10^{-16}
	60 days	EM	0.9519	0.9061	2.5×10^{-17}
		cEM	0.9492	0.9010	4.9×10^{-17}
Model-II	30 days	EM	0.8723	0.7610	6.3×10^{-11}
		cEM	0.9747	0.9500	1.1×10^{-20}
	60 days	EM	0.9817	0.9637	2.3×10^{-22}
		cEM	0.9684	0.9378	1.6×10^{-19}

4.3.6 Comparison with Existing Methods

We have made a clear example of the advantages of the cEM analysis compared to the existing EM/Expa analysis by calculation speed and accuracy. The speed and accuracy of cEM and EM/Expa analyses for the metabolic models are shown in table 4.7. The EM/Expa analyses consist of two steps: EM/Expa extraction by CNA and emftool, and flux prediction by MEP. On the other hand, the cEM analysis consists of three steps: FBA, EM decomposition for cEM extraction, and flux prediction by MEP. The calculation speed of cEM analysis was 35-fold and 168-fold higher than the EM analysis by CNA, 4-fold and 6-fold higher than Expa analysis, and 22-fold and 30-fold higher than the efmtool for model-I and model-II, respectively, whereas they did hardly deteriorated the prediction accuracy (Badsha et al., 2014). Both the processes of EM extraction (or FBA+EM decomposition) and MEP-based flux prediction were remarkably accelerated.

4.3.7 Critical Numbers of cEMs

It is important to determine the critical cEMs critically responsible for flux prediction in order to guarantee the quality of the flux prediction. As shown in figures 4.3F, 4.4F, 4.5F, 4.6F, a small number of cEMs was enough for the accurate prediction. We can check the critical cEMs to guarantee the quality of the flux prediction and to validate the algorithm. Consequently speaking, users can use all cEMs without any problems. At the critical number of cEMs, the prediction difference between cEM and EM analyses turned to a very gradual decrease or to be saturated with respect to the number of cEMs. Note that the critical number is less than 2×(the flux number to be estimated by FBA). If a large number of critical cEMs are required, we recommend a random sampling method, Constraint-based Reconstruction and Analysis (COBRA) (Becker et al., 2007) that can produce more flux distributions than the employed FBA.

4.3.8 Application to a Large-scale/Genome-scale Metabolic Network

To demonstrate the feasibility of cEM analysis and its applicability to large-scale or genome-scale metabolic models, we applied it to model-III (Agren et al., 2012). The EM/Expa analyses by CNA and efmtool were unable to find any EMs/Expas due to of calculation complexity or memory limitations. By contrast, use of cEM analysis generated 804 cEMs using 2,249 flux distributions and estimated the flux distribution by MEP, as shown in table 4.7. The cEM analysis is found to be applicable to a genome-scale network. Note that no prediction error is calculated, because there is no experimental flux data (Badsha et al., 2014).

Table 4.7: Calculation speed and accuracy for the prediction of flux distributions by the cEM and EM/Expa analyses.

Model	Method		# EM	Total running time(s)	Prediction error
Model-I	EM	CNA	122126	600+780.871=1380.871[a]	0.0233
		efmtool	122126	95+790.628=885.628[a]	0.0255
	Expa		1215	65+85.405=150.405[a]	0.0294
	cEM		202	7.94+30+1.561=39.501[b]	0.0268
Model-II	EM	CNA	321416	6000+1000.568=7000.568[a]	0.0813
		efmtool	321416	180+1070.326=1250.326[a]	0.0737
	Expa		3045	89+163.965=252.965[a]	0.0836
	cEM		295	8.08+32+1.605=41.685[b]	0.0975
Model-III	EM	CNA	*	*	*
		efmtool	*	*	*
	Expa		*	*	*
	cEM		804	3517.2+7871.5+ 31.80 =11420.5[b]	No experimental data

[a] The EM/Expa analyses consist of two steps: EM/Expa extraction and flux prediction by MEP. [b] The cEM analysis consists of three steps: FBA, EM decomposition, and flux prediction by MEP. * Not applicable due to calculation complexity.

4.4 Conclusion

We have presented a method for decomposing EM by implementing the EM decomposition and flux distributions predicted by MEP method, which demonstrations a fast and efficient analysis for large-scale metabolic model. The FBA is used to determine many possible ranges of metabolic flux distributions that the EM decomposition method requires as input data. The MEP is used as an objective function for optimizing the coefficients of cEMs/EMs and avoid to the scalar product problem. To demonstrate the feasibility of cEM analysis, we predicted the flux distributions of one synthetic metabolic network, two metabolic networks of *E. coli* and a large-scale metabolic network of head and neck cancer cells. The cEM analysis optimized the flux distribution much faster than EM analysis without deteriorating the prediction accuracy (Table 4.7). The cEM analysis accurately predicts the metabolic flux distributions of metabolic networks of *E. coli* and it's applicable to large/genome–scale metabolic network model. The cEM analysis greatly reduced the

computational time and memory cost, enabling analysis of a genome-scale metabolic network. It is useful to plan a genetic engineering strategy for large-scale metabolic networks producing of useful compounds.

Chapter 5

Conclusion, Scope and Future Research Interest

5.1 Conclusion

Metabolism encompasses all life-sustaining biochemical processes and it is playing an essential role in various aspects of biology, including the development and progression of many serious diseases (Deerardinis and Thompson, 2012). However, metabolism is a highly complex of a living cell involves several thousands of small molecules and their conversion, a full analysis of such a metabolic network is only feasible using mathematical or computational approaches. In addition, metabolism differs significantly from cell to cell and over different contexts. Systemic approaches to the study of a biological cell or tissue rely increasingly on the use of context-specific metabolic network models. Biological data are important in medical area for specified purposes such as patient documentation, disease presentation, statistical documentation, etc. The integration of heterogeneous biological data and model building have become essential activities in biological research as technological advancements continue to empower the measurement of biological data of increasing diversity and scale.

The current central challenge in the development of systems biology is the integration of heterogeneous biological data to generate predictive computational models. Modeling and simulation of biochemical networks are invaluable tools used by researchers to investigate cellular behavior and help in the interpretation of data arising from quantitative experiments. Quantitative methods for modeling of biological networks require accurate knowledge of the biochemical reactions, their stoichiometric and kinetic parameters, and in the case of

metabolic pathway modeling, initial concentrations of metabolites and enzymes (Smallbone et al., 2013). In many cases, such experimentally derived parameters are unavailable. Therefore, metabolic network analysis becomes a core method for constructing a mathematical model that predicts the flux distribution which gives us a good idea of what is happening in an organism and how the organisms work under different external environmental conditions for large-scale metabolic networks.

The biochemical reactions which illustrate various portions of the metabolism are depicted using a metabolic network. Constraint-based metabolic network analysis has focused on two approaches, optimization-based and pathway-based analysis, which are used for predicting the steady-state intracellular metabolic fluxes from the metabolic network by integrating of the experimental data from genomics, transcriptomics, proteomics, metabolomics, and fluxomics, which are determined by high-throughput technologies. The FBA, rFBA, MOMA, ROOM, FVA, with some genetically modified algorithms, e.g., OptKnock, RobustKnock, OptReg, OptGene, OptGene, OptForce are used to integrate the heterogeneous biological data into the metabolic network by the optimization-based analysis methods. On the other hand, pathway-based analysis methods, e.g., MFA, EM, Expa, CEF, mCEF, ECF, GMF are used to integrate the heterogeneous biological data into metabolic network. Integrated biological network analysis is used by IOMA and iMAT to integrate quantitative proteomics and metabolomics data, and tissue-specific gene and protein expression data with the genome-scale metabolic network.

However, pathway-based analysis is the most widely used and more advantageous than optimization-based analysis, because it generally employs without the specifying the cellular objective function. Pathway-based analysis is to offer a great opportunity for studying functional and structural properties of metabolic pathways. Pathway-based analysis facilitates understanding or designing a complex metabolic system and enables prediction of steady-

state metabolic flux distributions by EM and Expa analyses. EM analysis is potentially effective in integrating transcriptome or proteome data into metabolic network analyses and minimal set of reactions that can operate in a steady state, while Expa analysis is a subset of EM that contains one additional constraint to make all Expas systematically independent. EMs/Expas are the building blocks of the metabolic network and it has numerous applications in chemical engineering and biochemistry for the study of phenotype of wild type and mutant cells under particular conditions. The EM coefficients (EMCs) indicate the quantitative contribution of their associated EMs and can be estimated by maximizing as a particular objective function.

The principal drawback of the ordinary EM/Expa analysis is that the number of EMs/ Expas in a metabolic network suffers from a combinatorial explosion. The computational time increases exponentially with an increase in network sizes, which demonstrate that the computation of the all EMs/Expas expensive and infeasible for large-scale networks. Another problem is rising for estimating the EMCs to predict the flux distribution due to no specific objective biological functions are available and EMs can be described by different scalar products of each EM, but the predicted fluxes must be independent of them. To overcome such an existing problem, in this thesis we present a new method to analyze complex metabolic networks, cEM analysis. It combines an EM decomposition method, using the FBA method, with the MEP to predict the metabolic flux distributions. The predicted flux distributions are compared with experimental data and statistically analyzed.

To demonstrate the feasibility of cEM analysis, we compared it with EM/Expa analysis by using an artificial model, two medium-scale metabolic networks of *E. coli* and a genome-scale metabolic network of head and neck cancer cells. The cEM analysis greatly reduces the number of EM, computational time, memory cost and exposing a new window for a large-scale metabolic network analysis. The computational timings and accuracy are presented in

table 4.7, which clearly show that the cEM method is faster than EM/Expa analyses by CNA and emftool. The predicted flux distribution by cEMs analysis is very consistent with that by the EM/Expa analyses, whereas they did hardly deteriorated the prediction accuracy. Application of cEM analysis to GMF accurately predicts the flux distributions of genetic mutants under particular conditions.

5.2 Scope of the Study

It is very informative and important to analyze the physiological state of microorganisms, e.g., cell growth and biosynthesis by metabolic flux distributions, while the flux distribution data are not available in human cells, compared with proteome and transcriptome data, due to experimental complexity. It is critically important to predict flux distributions from available transcriptome and proteome data and to characterize the physiological state of diseased cells. This study not only contributes to life science in terms of a complete understanding of biological systems, but also gives great influence to advanced biotechnology for analysis of genome-scale metabolic networks..

5.3 Future Research Interest

Considering the previous body of research and undertaken work, future research interest may consist of several directions. Some of the major ideas may be outlined as:

- Human metabolic network analysis by cEM algorithm.
- Producing of useful compounds from the large-scale metabolic networks by the cEM analysis.
- Predicts metabolic flux distributions of disease- or tissue- specific cells under particular conditions.

- To develop the disease diagnostic tools from the metabolic network of disease cells by the cEM analysis.

- Establishment of the theoretical validation of cEM analysis.

Acknowledgements

First of all, I give thanks to Almighty Allah who gave me the confidence, the wisdom and the means to follow this work to its conclusion and for giving me the strength to overcome all the obstacles in the completion of this research work.

I would like to thank all people I met at Kyushu Institute of Technology (KIT) and those who have made a comfortable environment, tremendous support, inspiration and encouragement extended to me during my graduate time, as well as expressing my apology that I could not mention personally one by one.

I owe to express my heartfelt thanks and deepest gratitude to Prof. Dr. Hiroyuki Kurata, my dear supervisor and mentor, for being ultimately supportive, understanding and helpful during my study in KIT. It has been an immense honor to be his Ph.D student and have him as a mentor, colleague and collaborator. Professor Kurata is one of the kindest, gentlest and smartest people I have ever known, and I found that I was very lucky to work with him and he provided me with the unique opportunity to pursue my PhD studies in his group. This thesis and all of my research papers would have not been possible without or with less support and guidance from my advisor. His immense knowledge, perpetual enthusiasm for science and research, and thoughtful insights have helped me get on the right tract for my research and kept inspiring me. As an excellent scientist with rich experience, he has not only complemented my academic knowledge, but also enlarged my vision and improved my research skills. At the same time, he was always there for me to consult and discuss with, or check out my papers even during his weekend or on vacation. I believe that if I leave KIT, I would miss his dedication and kindness so much.

Special thanks are due to Professor Dr. Junshi Sakamoto, Professor Dr. Tetsushi Yada, Associate Professor Dr. Shunsuke Aoki and Associate Professor Dr. Katsuya Nagayama, for

serving the examination committee. Their valuable comments, continuous inspiration, assistance and insightful suggestions have greatly improved my defense presentation and also my thesis work.

I would like to express my profound appreciation to people in my lovely laboratory who were always there to help me to complete my research, especially Dr. Kazuhiro Maeda, Dr. Yu Matsuoka, Mr. Ryo Tsuboi, Mr. Yuta Matsumoto and Mr. Daisuke Koishi who helped me a lot with software setups, translation of paperwork, and chatting in Japanese.

I wish to extend my sincere gratitude to the Grant-in-Aid for Scientific Research (B) (25280107) from the Japan Society for the Promotion of Science and Ministry of Economy, Trade and Industry (METI), Japan and 100th Anniversary Memorial Scholarship of KIT for their financial support during my study in Japan. Without these important resources, I would not have totally focused on completion of my research.

I am greatly indebted to my wife, Nusrat Jahan. She has sacrificed many weekends due to my research work, but even though she insisted encouraged me. I am grateful to her as she believed in me, helped me, and supported me.

Finally, I wish to dedicate my gratefulness to my beloved parents, friends and well-wishers, for their patience and encouragement throughout my studies to the completion of this thesis.

Md. Bahadur Badsha
Kyushu Institute of Technology, Japan.

References

[1] Acuna, V, Chierichettia, F, Lacroixb, V, Spaccamelaa, AM, Sagotb, MF and Stougie, L: Modes and cuts in metabolic networks: complexity and algorithms, *Biosystems*, **95**, 51–60, 2009.

[2] Agren, R, Bordel, S, Mardinoglu, A, Pornputtapong N, Nookaew, I and Nielsen, J: Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT, *PLoS Comput. Biol*., **8** (5), e1002518, 2012.

[3] Ahmed, Z: Physical Biology: From Atoms to Medicine, *Imperial College Press*.p. 339, 2008.

[4] Badsha, MB, Jahan, N, Mollah, MNH and Kurata, H: Metabolic Engineering for Systematic Organization and Analysis of Complex Metabolic Networks, *International Conference on Statistical Data Mining for Bioinformatics, Health, Agriculture and Environment,* Bangladesh, 21-24 December, 2012.

[5] Badsha, MB, Tsuboi, R and Kurata, H: Complementary elementary mode analysis for large-scale metabolic networks, *IPSJ SIG technical report*, BIO-35 (**5**), 1–2, 2013.

[6] Badsha, MB, Tsuboi, R and Kurata, H: Complementary elementary modes for fast and efficient analysis of metabolic networks, *Biochem. Eng. J.*, **90**, 121–130, 2014.

[7] Becker, SA and Palsson, BO: Genome-scale reconstruction of the metabolic network in staphylococcus aureus n315: an initial draft to the two-dimensional annotation, *BMC Microbiology*, **5**, 2005.

[8] Becker, SA, Feist AM, Mo, ML, Hannum, G, Palsson, BO and Herrgard, MJ: Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox, *Nat. Protoc*., 2, 727–738, 2007.

[9] Blazier, AS and Papin, JA: Integration of expression data in genome-scale metabolic network reconstructions, *Front Physiol*, **3**, 299, 2012.

[10] Borodina, I and Nielsen, J: From genomes to in silico cells via metabolic networks, *Curr. Opin. Biotechnol*, **16**, 350-355, 2005.

[11] Bordbar, A, Monk, JM, King, ZA and Palsson, BO: Constraint-based models predict metabolic and associated cellular functions, *Nature Reviews Genetics*, **15**, 107–120, 2014.

[12] Burgard, AP, Pharkya, P and Maranas, CD: OptKnock: a bi-level programming framework for identifying gene knockout strategies for microbial strain optimization, *Biotechnology and Bioengineering*, **84**(6), 647–657, 2003.

[13] Cakir, T, Kirdar, B and Ulgen, KO: Metabolic pathway analysis of yeast strengthens the bridge between transcriptornics and metabolic networks, *Biotechnol.Bioeng*., **86**, 251–260, 2004.

[14] Cakir, J, Kirdar, B, Onsan, ZI, Ulgen, KO and Nielsen, J: Effect of carbon source perturbations on transcriptional regulation of metabolic fluxes in *Saccharomyces cerevisiae*, *BMC Systems Biology*, **1**(18), 2007.

[15] Cakir, T, Kirdar, B, Onsan, Z, Ulgen, KO and Nielsen J: Effect of carbon source perturbations on transcriptional regulation of metabolic fluxes in *Saccharomyces cerevisiae*, *BMC Syst. Biol*., **1**, 18, 2007.

[16] Caspi, R, Foerster, H, Fulcher, CA, Hopkinson, R, Ingraham, H, Kaipa, P, Krummenacker, M, Paley, S, Pick, J, Rhee, SY, Tissier, C, Zhang, P and Karp, PD: MetaCyc: a multiorganism database of metabolic pathways and enzymes, *Nucleic Acids Research*, **34**, 511–516, 2006.

[17] Caspi, R, Foerster, H, Fulcher, A, Kaipa, P, Krummenacker, M, Latendresse, M, Paley, S, Rhee, SY, Shearer, G, Tissier, C, Walk, TC, hang, P and Karp, PD: The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases, *Nucleic Acids Research*, **36**, 623–631, 2008.

[18] Carlson, R and Srienc, F: Fundamental Escherichia coli biochemical pathways for biomass and energy production: identification of reactions, *Biotechnol. Bioeng*., **85** (1), 1–19 2004.

[19] Capra, JA and Singh, M: Predicting functionally important residues from sequence conservation, *Bioinformatics*, **23**, 1875–1882, 2007.

[20] Clarke, B: Stoichiometric network analysis, *Cell Biophysics*, **12**, 237–53, 1988.

[21] Croes, D, Couche, F, Wodak, SJ and Helde, JV: Metabolic PathFinding: inferring relevant pathways in biochemical networks, *Nucleic Acids Res*., **33**, 326–330, 2005.

[22] Covert, MW, Schilling, CH and Palsson, BO: Regulation of gene expression in flux balance models of metabolism, *J. Thor. Biol*., **213**, 73–88, 2001.

[23] Covert, M and Palsson, BO: Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*, *The Journal of Biological Chemistry*, **27**, 28058–28064, 2002.

[24] Covert, MW, Famili, I and Palsson, BO: Identifying constraints that govern cell behavior: A key to converting conceptual to computational models in biology, *Biotechnol. Bioeng*., **84**(7), 309–325, 2003.

[25] Covert, M and Palsson, BO: Constraints-based models: Regulation of gene expression reduces the steady-state solution space, *The Journal of Biological Chemistry*, **221**, 309–325, 2003.

[26] DeBerardinis, RJ and Thompson, CB:Cellular metabolism and disease: what do metabolic outliers teach us? *Cell*, 148, 1132–1144, 2012.

[27] Dellomonaco, C: Engineered Reversal of the beta oxidation cycle for the Synthesis of Fuels and Chemicals, *Nature*, **476**, 355-359, 2011.

[28] de Figueiredo, LF, Podhorski, A, Rubio, A, Kaleta, C, Schuster, S and Planes, FJ: Computing the shortest elementary flux modes in genome scale metabolic networks, *Bioinformatics*, **25**, 3158–3165, 2009.

[29] de Figueiredo, LF, Podhorski, A, Rubio, A, Kaleta, C, Beasley, JE, Schuster, S and Planes, FJ: Computing the shortest elementary flux modes in genome-scale metabolic networks, *Bioinformatics*, **25**(23), 3158–3165, 2009.

[30] Duarte, NC, Herrgard, MJ and Palsson, BO: Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome scale metabolic model, *Genome Research*, **14**(7), 1298–309, 2004.

[31] Duarte, NC, Becker, SA, Jamshidi, N, Thiele, I, Mo, ML, Vo, TD, Srivas, R and Palsson BO: Global reconstruction of the human metabolic network based on genomic and bibliomic data, *Proc. Natl. Acad. Sci. U.S.A*., **104**(6), 1777–1782, 2007.

[32] Edwards, JS and Palsson, BO: Metabolic flux balance analysis and the in silico analysis of Escherichia coli k-12 gene deletions, *BMC Bioinformatics*, **1**,1, 2000.

[33] Fell, DA and Small, JR: Fat synthesis in adipose tissue. an examination of stoichiometric constraints, *Biochemical Journal*, **238**, 781–786, 1986.

[34] Feist, AM, Herrgard, MJ, Thiele, I, Reed, JL and Palsson, BO: Reconstruction of biochemical networks in microorganisms, *Nature Reviews Microbiology*, **7**, 129–143, 2009.

[35] Feist, AM and Palsson, BO: The biomass objective function, *Curr.Opin.Microbiol,* **13** (3), 344–349, 2010.

[36] Forster, J, Famili, I, Fu, P and Palsson, BO: Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network, *Genome Research*, **13**, 644–653, 2003.

[37] Gagneur, J and Klamt, S: Computation of elementary modes: a unifying framework and the new binary approach, *BMC Bioinformatics*, **5**(175), 2004.

[38] Gayen, K and Venkatesh, KV: Analysis of optimal phenotypic space using elementary modes as applied to Corynebacteriumglutamicum, *BMC Bioinformatics*, **7**, 445, 2006.

[39] Goodman, SN: Toward Evidence-Based Medical Statistics: The P Value Fallacy, *Annals of Internal Medicine,* **130**, 995–1004, 1999.

[40] Green, ML, Kaiser, D, Krummenacker, M, Romero, P, Wagg, J and Karp, PD: Computational prediction of human metabolic pathways from the complete human genome, *Genome Biology*, **6**(1), 2004.

[41] Gudmundsson, S and Thiele, I: Computationally efficient flux variability analysis, *BMC Bioinformatics*, **11**, 489, 2010.

[42] Hatzimanikatis, V, Li, C, Ionita, JA, and Broadbelt, LJ: Metabolic networks: enzyme function and metabolite structure, *Current Opinion in Structural Biology*, *14*, 300–306, 2004.

[43] Hasbun, JE: Classical Mechanics with MATLAB Applications, *Jones and Bartlett, Sudbury*, 2008.

[44] Haus, UU, Klamt, S and Stephen, T: Computing knockout strategies in metabolic networks, *J. Comput. Biol*., **15**, 259–268, 2008.

[45] Hua, Q, Joyce, AR, Fong, SS, Palsson, BO: Metabolic analysis of adaptive evolution for in silico-designed lactate-producing strains, Biotechnol. Bioeng., 95, 992–1002, 2006.

[46] Ideker, T, Thorsson V, Ranish, JA, Christmas, R, Buhler, J, Eng, JK, Bumganer, R, Goodlett, DR, Aebersold, R and Hood, L: Integrated genomic and proteomic analyses of a systematically perturbed metabolic network, *Science*, **292**, 929–934, 2001.

[47] Ip, K, Colijn, C and Lun, DS: Analysis of complex metabolic behavior through pathway decomposition, *BMC Syst. Biol*., **5**, 91, 2011.

[48] Jiang, D: Network-based Metabolic Flux and Structure Analysis, 2006.

[49] Jevremovic, D, Trinh, CT, Srienc, F, Sosa, CP and Boley, D: Parallelization of null space algorithm for the computation of metabolic pathways, *Parallel Comput*., **37** (6–7), 261- 278, 2011.

[50] Jevremovic, D and Boley, D: Parallel computation of elementary flux modes in metabolic networks using global array, *The 6th IEEE Conference on Partitioned Global Address Space Programming Models*, Santa Barbara, CA, 2012.

[51] Jevremovic, D: Scalable Computation and Analysis of Elementary Flux Modes in Metabolic Networks, *PhD Thesis*, 2013.

[52] Jungreuthmayer, C, Ruckerbauer, DE and Zanghellini, J: regEfmtool: speeding up elementary flux mode calculation using transcriptional regulatory rules in the form of three state logic, *BioSystems*, **113**, 37–39, 2013.

[53] Kanehisa, M and Goto, S: KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Research*, **28**, 27–30, 2012.

[54] Kanehisa, M, Goto, S, Sato, Y, Furumichi, M and Tanabe, M: KEGG for integration and interpretation of large-scale molecular datasets, *Nucleic Acids Research*, **40**, D109– D114, 2012.

[55] Kauffman, KJ, Prakasha, P and Edwards, JS: Advances in flux balance analysis, *Current Opinion in Biotechnology*, **14**(5), 491–496, 2003.

[56] Kitano, H: Computational systems biology, *Nature*, **420**(6912), 206-210, 2002a.

[57] Kitano, H: Systems biology: a brief overview, *Science*, **295**(5560), 1662-1664, 2002b.

[58] Keseler, IM, Collado-Vides, J, Santos-Zavaleta, A, Peralta-Gil, M, Gama-Castro, S, Muniz-Rascado, S, Bonavides-Martinez, C, Paley, S, Krummenacker, M, Altman, T, Kaipa, P, Spaulding, A, Pacheco, J, Latendresse, M, Fulcher, C, Sarker, M, Shearer, AG, Mackie, A, Paulsen, I, Gunsalus, RP and Karp, PD: EcoCyc: a comprehensive database of *Escherichia coli* biology, *Nucleic Acids Research*, **39**, D583–590, 2011.

[59] Klipp, E, Liebermeister, W, Wierling, C, Kowald, A, Lehrach, H and Herwig, R: Systems Biology, *John Wiley and Sons*, first edition, 2009.

[60] Klamt, S, Rodriguez, JS and Gilles, ED: Structural and functional analysis of cellular networks with CellNetAnalyzer, *BMC Syst. Biol*., **1**, 2, 2007.

[61] Kim, J and Reed, J: OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains, *BMC Systems Biology*, **4**(53), 53–71, 2010.

[62] King, RD, Garrett, SM and Coghill, GM: On the use of qualitative reasoning to simulate and identify metabolic pathways, *Bioinformatics*, **21**, 2017–2026, 2005.

[63] Kurata, H, Zaho, Q, Okuda, R and Shimizu, K: Integration of enzyme activities into metabolic flux distributions by elementary mode analysis, *BMC Syst. Biol*., **1**, 31, 2007.

[64] Lacroix, V, Cottret, L, Thebault, P and Sagot, M: An introduction to metabolic network analysis and their structural analysis, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **5**(4), 594–617, 2008.

[65] Lee, JM, Gianchandani, EP and Papin, JA: Flux balance analysis in the era of metabolomics, *Briefings in Bioinformatics*, **7**(2), 140–150, 2006.

[66] Lezon, TR, Banavar, JR, Cieplak, M, Maritan, A and Fedoroff, NV: Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns, *Proc. Natl. Acad. Sci. U. S. A*., **103**, 19033–19038, 2006.

[67] Luenberger, DG: Linear and nonlinear programming, *Springer*, second edition, 2003.

[68] Ma, H and Zenf, AP: Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms, *Bioinformatics*, **19**(2), 270–277, 2003.

[69] Ma, H, Sorokin, A, Mazein A, Selkov, A, Selkov, E, Demin, O and Goryanin, I: The Edin-burgh human metabolic network reconstruction and its functional analysis, *Mol. Syst. Biol*., **3** 2007.

[70] Machado, D, Soons, Z, Patil, KR, Ferreira, EC and Rocha, I: Random sampling of elementary flux modes in large-scale metabolic networks, *Bioinform. ECCB*, **28**, i515–i521, 2012.

[71] Mahadevan, R and Schilling CH: The effects of alternate optimal solutions in constraint-based genome-scale metabolic models, *Metabolic Engineering*, **5**(4), 264–276, 2003.

[72] Marashi, SA: Constraint-based Analysis of Substructures of Metabolic Networks, *PhD thesis Dissertation, Berlin, Germany*, 2011.

[73] Martin, LC, Gloor, GB, Dunn, SD and Wahl, LM: Using information theory to search for co-evolving residues in proteins, *Bioinformatics*, **21**,4116–4124, 2005.

[74] Milne, CB, Kim, PJ, Eddy, JA and Price, ND: Accomplishments in genome-scale insilico modeling for industrial and medical biotechnology, *Biotechnol. J*., **4**, 1653–1670, 2009.

[75] Orth, JD, Thiele, I and Palsson, BO: What is flux balance analysis? *Nature Biotechnology*, **28**(3), 245-248, 2010.

[76] Papin, JA, Stelling, J, Price, ND, Klamt, S, Schuster, S and Palsson, BO: Comparison of network-based pathway analysis methods, *Trends Biotechnol*., **22**(8), 400–405, 2004.

[77] Papin, JA, Stelling, J, Price, ND, Klamt, S, Schuster, S and Palsson, BO: Comparison of network-based pathway analysis methods, *Trends Biotechnol*., **22**, 400–405, 2004.

[78] Patil, KR and Nielsen, J: Uncovering transcriptional regulation of metabolism by using metabolic network topology, *Proc.Natl.Acad.Sci.U.S.A.,* 102(8), 2685–2689, 2005.

[79] Patil, KR, Rocha, I, Forster, J and Nielsen, J: Evolutionary programming as a platform for in silico metabolic engineering, *BMC Bioinformatics*, **6**(1), 308, 2005.

[80] Park, JM, Kim, TY, and Lee, SY: Constraints-based genome-scale metabolic simulation for systems metabolic engineering, *Biotechnology Advances*, **27**, 979–988, 2009.

[81] Patnaik, R and Liao, J: Engineering of Escherichia coli central metabolism for aromatic metabolite production with near theoretical yield, *Appl. Environ. Microbiol.*60(11), 3903-3908, 1994.

[82] Penrose, R: A generalized inverse for matrices, *Proc. Cambridge Phil. Soc*, **51**, 406–413, 1955.

[83] Pearson, K: Note on regression and inheritance in the case of two parents, *Proceedings of the Royal Society of London*, **58** : 240–242, 1895.

[84] Pharkya, P, Burgard, AP and Maranas, CD: Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock, *Biotechnol. Bioeng*., **84**(7), 887–99, 2003.

[85] Pharkya, P and Maranas, CD: An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems, *Metabolic Engineering*, **6**(8), 1–13, 2006.

[86] Price, ND, Reed, JL and Palsson, BO: Genome-scale models of microbial cells: evaluating the consequences of constraints, *Nat. Rev. Microbiol*., **2**, 886–897, 2004.

[87] Price, ND, Thiele, I and Palsson, BO: Candidate states of Helicobacter pylori's genome-scale metabolic network upon application of "loop law" thermodynamic constraints, *Biophys. J*., 90, 3919–3928, 2006.

[88] Raman, K and Chandra, N: Flux balance analysis of biological systems: applications and challenges, *Briefings in Bioinformatics*, **10**(4), 435–449, 2009.

[89] Ranganathan, S, Suthers, PF and Maranas, CD: Optforce: An optimization procedure for identifying all genetic manipulations leading to targeted overproductions, *PLoS Comput. Biol*., 6(4), 2010.

[90] Schilling, CH and Palsson, BO: Assessment of the Metabolic Capabilities of Haemophilus Influenzae Rd through a Genome-scale Pathway Analysis, *J. Theor. Biol*., **203**(3), 249–283, 2000.

[91] Schilling, CH, Schuster, S, Palsson, BO and Heinrich, R: Metabolic pathway analysis: Basic concepts and scientific applications in the post-genomic era, *Biotechnol.Prog*., **15**, 296–303, 1999.

[92] Schilling, CH, Edwards, JS and Palsson, BO: Towards metabolic phenomics: Analysis of genomic data using flux balances, *Biotechnol.Prog*., **15**(3), 288–295, 1999.

[93] Schilling, CH, Letscher, D and Palsson, BO: Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective, *J. Theor. Biol*., **203**, 229–248, 2000.

[94] Schilling, CH, Covert, MW, Famili, I, Church, GM, Edwards, JS, and Palsson, BO: Genome-scale metabolic model of helicobacter pylori 26695, J. Bacteriol, **184**(16), 4582–4593, 2002.

[95] Schuster, S and Hilgetag, C: On elementary flux modes in biochemical reaction systems at steady state, *J. Biol. Syst*., **2**, 165{182, 1994.

[96] Schuster, S, Dandekar, T and Fell, DA: Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering, *Trends Biotechnol*., **17**, 53–60, 1999.

[97] Schwender, J, Goffman, F, Ohlrogge, JB and Hill, YS: Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds, *Nature*, **432**, 779–782, 2004.

[98] Schwartz, JM and Kanehisa, M: A quadratic programming approach for decomposing steady-state metabolic flux distributions onto elementary modes, *Bioinformatics*, 21 (Suppl. 2), ii 204–ii 205, 2005.

[99] Schwartz, JM and Kanehisa, M: Quantitative elementary mode analysis of metabolic pathways: the example of yeast glycolysis, *BMC Bioinformatics*, **7**, 186, 2006.

[100] Segre, D, Vitkup, D and Church, GM: Analysis of optimality in natural and perturbed metabolic networks, *Proc. Natl. Acad. Sci. U.S.A*., **99**(29), 15112–15117, 2002.

[101] Shlomi, T, Berkman, O and Ruppin, E: Regulatory on/off minimization of metabolic flux changes after genetic perturbations, *Proc. Natl. Acad. Sci. U.S.A*., **102**(21), 7695–7700, 2005.

[102] Shlomi, T, Eisenberg, Y, Sharan, R and Ruppin, E: A genome-scale computational study of the interplay between transcriptional regulation and metabolism, *Molecular Systems Biology*, **3**(101), 2007.

[103] Shlomi, T, Cabili, MN, Herrgard, MJ, Palsson, BO and Ruppin, E: Network-based prediction of human tissue-specific metabolism, *Nature Biotechnology*, **26**(9), 2008.

[104] Shannon, CE; A mathematical theory of communication, *Bell Syst. Technol. J*., **27**, 379–423, 623–656, 1948.

[105] Siddiquee, KAZ, Arauzo-Bravo, MJ and Shimizu, K: Effect of a pyruvate kinase (pykF-gene) knockout mutation on the control of gene expression and metabolic fluxes in *Escherichia coli*, *FEMS Microbiol. Lett*., **235**, 25–33, 2004.

[106] Smallbone, K, Messiha, HL, Carroll, KM, Winder, CL, Malys, N, Dunn, WB, Murabito, E, Swainston, N, Dada, JO, Khan, F, Pir, P, Simeonidis, E, Spasić, I, Wishart, J, Weichart, D, Hayes, NW, Jameson, D, Broomhead, DS, Oliver, SG, Gaskell, SJ, McCarthy, JE, Paton, NW, Westerhoff, HV, Kell, DB and Mendes P: A model of yeast glycolysis based on a consistent kinetic characterisation of all its enzymes, *FEBS Lett*., **587**, 2832–2841, 2013.

[107] Stephanopoulos, GN, Aristidou, AA and Nielsen, J: Metabolic Engineering: Principles and Methodologies, *San Diego: Academic Press*, 1998.

[108] Steel, RGD and Torrie, JH: Principles and Procedures of Statistics with Special Reference to the Biological Sciences, *McGraw Hill, New York*, 1960.

[109] Stelling, J, Klamt, S, Bettenbrock, K, Schuster, S and Gilles, ED: Metabolic network structure determines key aspects of functionality and regulation, *Nature*, **420**,190–193, 2002.

[110] Suthers, P, Dasika, M, Kumar, V, Denisov, G, Glass, J and Maranas, C: A genome scale metabolic reconstruction of Mycoplasma genitalium iPS189, *PLoS Computational Biology*, **5**(2), 2009.

[111] Terzer, M and Stelling, J: Large-scale computation of elementary flux modes with bit pattern trees, *Bioinformatics*, **24**, 2229–2235, 2008.

[112] Tepper, N and Shlomi, T: Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways, *Bioinformatics*, **26**(4), 536– 543, 2009.

[113] Thiele, I, Vo, TD, Price, ND and Palsson, BO: Expanded metabolic reconstruction of helicobacter pylori (iit341 gsm/gpr): an in silico genome-scale characterization of single- and double-deletion mutants, *J. Bacteriol*, **187**(16), 5818–5830, 2005.

[114] Thiele, I et al: A community-driven global reconstruction of human metabolism, *Nature Biotechnology*, **31**(5), 419-425, 2013.

[115] Toya, Y, Ishii, N, Nakahigashi, K, Hirasawa, T, Soga, T, Tomita, M and Shimizu, K: ^{13}C-metabolic flux analysis for batch culture of Escherichia coli and its Pyk and Pgi gene knockout mutants based on mass isotopomer distribution of intracellular metabolites. *Biotechnol.Prog.*, **26** (4), 975–992, 2010.

[116] Trinh, CT, Wlaschin, A and Srienc, F: Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism, *Appl. Microbiol. Biotechnol*., **81**, 813–826, (2009).

[117] Van, GWM and Heijnen, JJ: A metabolic network stoichiometry analysis of microbial growth and product formation. *Biotechnol. Bioeng*, **48** (6), 681–698, 1995.

[118] Varma, A and Palsson, BO: Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use, *Nature Biotechnology*, **12**, 994–998, 1994a.

[119] Varma, A and Palsson, BO: Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110, *Appl.nviron. Microbiol*, **60** (10), 3724–3731, 1994b.

[120] Voit, E and Torres, NV: Pathways Analysis and Optimization in Metabolic Engineering, *Cambridge:University Press*, 2002.

[121] Waters, KM, Liu, T, Quesenberry, RD, Willse, AR, Bandyopadhyay, S, Kathmann, LE, Weber, TJ, Smith, RD, Wiley, HS and Thrall, BD: Network analysis of epidermal growth factor signaling using integrated genomic, proteomic and phosphorylation data, *PLoS ONE*, **7**, 3, 2012.

[122] Wang, S: Metabolic Network, *University of York, UK*, 2011.

[123] Wang, Z and Zhang, J: Abundant Indispensable Redundancies in Cellular Metabolic Networks,*Genome Biol. Evol.*, 1, 23-33, 2009.

[124] Wiback, SJ, Mahadevan, R and Palsson, BO: Using metabolic flux data to further constrain the metabolic solution space and predict internal flux patterns: the Escherichia coli spectrum, *Biotechnol. Bioeng*., **86**, 317–331, 2004.

[125] Yizhak, K, Benyamini, T, Liebermeister, W and Ruppin, E: Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model, *Bioinformatics*, **26**(12), 255–260, 2010.

[126] Zaho, Q and Kurata, H: Maximum entropy decomposition of flux distribution at steady state to elementary modes, *J. Biosci. Bioeng*., **107**, 84–89, 2009a.

[127] Zaho, Q and Kurata, H: Genetic modification of flux for flux prediction of mutants, *Bioinformatics*, **25**, 1702–1708, 2009b.

[128] Zaho, Q and Kurata, H: Use of maximum entropy principle with Lagrange multipliers extends the feasibility of elementary mode analysis, *J. Biosci. Bioeng*., **110** (2), 254–261, 2010.

[129] Zhao, Y, Tamura, T, Akutsu, T and Vert, JP: Flux balance impact degree: a new definition of impact degree to properly treat reversible reactions in metabolic networks, *Bioinformatics*, **29** (17), 2178–2185, 2013.

[130] Zomorrodi, AR, Suthers, PF, Ranganathan, S and Maranas, CD: Mathematical optimization applications in metabolic networks, *Metabolic Engineering*, **12**, 672–686, 2012.

[131] Zur, H, Ruppin, E and Shlomi, T: imat: an integrative metabolic analysis tool, *Bioinformatics*, **26**(24), 3140–3142, 2010.

Abbreviations and Symbols

Abbreviations

ATP : Adenosine triphosphate

CEF : Control effective flux

cEM :Complementary elementary mode

CNA : CellNetAnalyzer

COBRA : Constraint-based reconstruction and analysis

DNA : Deoxyribonucleic acid

E. coli : Escherichia coli

efmtool : Elementary flux mode tool

EM : Elementary mode

EMC : Elementary mode coefficient

ECF : Enzyme control flux

EM Decomp : Elementary mode decomposition

Expa : Extreme pathway

ECFLP : Enzyme control flux linear programing

FVA : Flux variability analysis

FBA : Flux balance analysis

GER : Gene Enzyme Reaction

GMF : Genetic modification of flux

IOMA : Integrative Omics-Metabolic Analysis

iMAT : Integrative Metabolic Analysis Tool

KEGG : Kyoto Encyclopedia of Genes and Genomes

LP : Linear programming

mCEF : Modified control effective flux

MEP : Maximum entropy principle

MILP : Mixed-integer linear programming

MFA : Metabolic flux analysis

MM : Michaelis-Menten

MOMA : Minimization of metabolic adjustment

QP : Quadratic programming

rFBA : Regulatory FBA

ROOM : Regulatory on/off minimization

SR-FBA : Steady state Regulatory FBA

Symbols

S : Stoichiometric matrix

v : Flux vector

i : Number of flux

n : Total number of flux

v_j : *j*-th number of flux

l : Number of metabolites

m : Total number of metabolites

C_l : Concentration of the *l*-th metabolite

ex : Extracellular metabolite

R : Null space matrix

wt : Wild-type

mut : Mutant-type

d : Degrees of freedom

k : Number of constraints

nd : Number of the measurable /determined fluxes

Z : Objective function

c_i : Weight coefficient for the *i*-th flux

lb : Lower boundaries

ub : Upper boundaries

KO : Indices of deleted reactions

γ : Specified parameter

ζ : Specified parameter

y_i : Binary variable

η : Specified parameter

ϕ : Specified parameter

$\mathbf{P}$: Elementary mode matrix

$\boldsymbol{\lambda}$: Elementary mode coefficient

$\mathbf{e}_i$: ith EM vector

$P_{i,j}$: Normalized element of the i-th reaction in the j-th EM

$\gamma_{j,\mathrm{SPEOBJ}}$: Ratio of EM output for the specific biological function

CEF_i : ECF of i-th reaction

$P_{\mathrm{SPEOBJ}}^{\max}$: Maximum element in the row of biological function

$\gamma_{j,\mathrm{SPEOBJ}}^{m}$: Efficiency of the j-th EM of a genetic mutant for the specific biological function

EAP_i : Relative gene expression for i-th reaction

π_i : Correction factor for i-th reaction

$\mathrm{ge}_{i,j}$: Gene expression for i-th reaction in the j-th EM

$\mathrm{mECF}_i(mut)$: mECF of the i-th reaction of a mutant type

$\mathrm{mECF}_i(wt)$: mECF of the i-th reaction of a wild type

$\theta_i(wt, mut)$: Relative change of a mutant to wild type of mECF

λ^{wt} : EMC of Wild-type

λ^{mut} : EMC of mutant-type

$a_{i,j}$: Relative enzyme activity

β : Factor parameter

v^{mut} : Flux distribution of the mutant

R_H : Highly expressed

R_L : Lowly expressed

v_i^{irrev} : *i*-th irreversible flux

v_i^{rev} : *i*-th reversible flux

$\boldsymbol{p}^{(k)}$: Set of cEMs

K : The number of cEMs

$\boldsymbol{\lambda}_k$: coefficient of *k*-th cEM

$\mathbf{v}_d$: Flux vector with *d*-th determined reactions

$\mathbf{P}_d$: Sub-matrix of EM/Expa/cEM matrix

ρ_j : Probability of *j*-th EM/Expa/cEM

ne : Total number of EM/Expa/cEM

$p_{substrate\ uptake,j}$: Element of the *j*-th EM/Expa/cEM

$v_{substrate\ uptake}$: Flux for substrate uptake

v_r : *r*-th determined flux

$\mathbf{x}_{r,j}$: Converted matrix from EM/Expa/cEM matrix $\mathbf{P}_d$

$p_{r,j}$: Element of the *r*-th determined flux and *j*-th EM/Expa/cEM

$\boldsymbol{\lambda}_i$: Coefficient vector of *i*-th cEMs/EMs

ψ : Nonlinear parameter

$\mathbf{v}^{target}$: Flux distribution of the target model

λ^{target} : EMC of the target model

numuptake : Row index that corresponds to the uptake or input flux

$v_{i,prediction}$: Predicted flux for the *i*-th reaction

$v_{i,exp}$: Experimental data of the *i*-th reaction

$v_{g,cEM}$: predicted fluxes for the *g*-th reaction by cEM

$v_{g,\ EM/Expa}$: Predicted fluxes for the *g*-th reaction by EM/Expa

nu : Number of the unmeasured fluxes.

Lists of Tables

Lists of Figures

Appendix A.

Metabolic Network Models

A.1 Escherichia coli (*E. coli*)

Table A.1 Metabolic reaction list for *E. coli.*

No	Gene	Reaction
1	Biomass formation	0.14176 Glyc3P + 26.2949 ATP + 0.60097 Ala + 0.10124 Cys + 0.26647Asp + 0.30747 Glu + 0.2048 Phe + 0.67725 Gly + 0.10473 His + 0.32116 Ile+ 0.37935 Lys + 0.49804 Leu + 0.16989 Met + 0.26647 Asn + 0.24436 Pro+ 0.29091 Gln + 0.32698 Arg + 0.38031 Ser + 0.28044 Thr + 0.46778 Val+ 0.062835 Trp + 0.15244 Tyr + 0.1489 rATP + 0.18319rGTP + 0.11366 rCTP+ 0.12273 rUTP + 0.023904 dATP + 0.024582 dGTP + 0.024582 dCTP + 0.023904 dTTP + 0.28352 avg_FS + 0.0069264 UDPGlc+0.010368 CDPEth+0.010368 OH_myr_ac + 0.010368C14_0_FS +0.010368 CMP_KDO + 0.010368 NDPHep+ 0.0069264 TDPGlcs + 0.1656 UDP_NAG + 0.01656 UDP_NAM + 0.01656 di_am_pim + 0.0924 ADPGlc ==> Biomass
2	Nitrogen uptake	==> N
3	CO2 exchange	CO2 <==>
4	Sulfur uptake	4 ATP + 4 NADPH ==> S
5	*pts*	PEP + GLC ==> G6P + Pyr
6	*glk*	ATP + GLC ==> G6P
7	Succinate exchan	Succ ==>
8	*gps*	DHAP + NADH <==> Glyc3P
9	Lactate exchange	Lac ==>
10	Ethanol exchange	Eth ==>
11	Acetate exchange	Ac ==>
12	Formate exchange	Form ==>
13	*pgi*	G6P <==> F6P
14	*fbp*	F16P ==> F6P
15	*fba*	F16P <==> DHAP + G3P
16	*tpi*	DHAP <==> G3P
17	*gap*	G3P <==> DPG + NADH
18	*pgk*	DPG <==> 3PG + ATP
19	*Gpm*	3PG <==> 2PG

Table A.1 (continued)

20	*eno*	2PG <==> PEP
21	*pyk*	PEP ==> Pyr + ATP
22	*pps*	Pyr + 2 ATP ==> PEP
23	*lpd*	Pyr ==> AcCoA + NADH + CO2
24	*glt*	AcCoA + OxA ==> Cit
25	*acn*	Cit <==> ICit
26	*icd*	ICit <==> alKG + NADPH + CO2
27	*sucAB*	alKG ==> SuccCoA + NADH + CO2
28	*sucCD*	SuccCoA <==> Succ + ATP
29	*sdh*	Succ ==> Fum + QuiH2
30	*frd*	Fum + QuiH2 ==> Succ
31	*Fum*	Fum <==> Mal
32	*mdh*	Mal <==> OxA + NADH
33	*aceA*	ICit ==> Succ + Glyox
34	*aceB*	AcCoA + Glyox ==> Mal
35	*zwf*	G6P <==> PGlac + NADPH
36	*adhE*	AcCoA + NADH <==> Adh
37	*adhE*	NADH + Adh <==> Eth
38	*pgl*	PGlac ==> PGluc
39	*gnd*	PGluc ==> Rl5P + NADPH + CO2
40	*rpe*	Rl5P <==> X5P
41	*rpi*	Rl5P <==> R5P
42	*tktAB*	R5P + X5P <==> G3P + S7P
43	*tal*	G3P + S7P <==> F6P + E4P
44	*tktAB*	E4P + X5P <==> F6P + G3P
45	*edd*	PGluc ==> KetoPGluc
46	*eda*	KetoPGluc <==> G3P + Pyr
47	*pck*	OxA + ATP ==> PEP + CO2
48	*ppc*	PEP + CO2 ==> OxA
49	*pta*	AcCoA <==> AcP
50	*ack*	AcP <==> ATP + Ac
51	*pfl*	Pyr ==> AcCoA + Form
52	*ldh*	Pyr + NADH <==> Lac
53	*nuo*	NADH <==> QuiH2 + 2 H_ex
54	*pntA*	NADH + H_ex <==> NADPH
55	ATP Synthesis	3 H_ex <==> ATP
56	ATPdrain	ATP ==>
57	*aro*	2 PEP + E4P + ATP + NADPH ==> Chor
58	*prsA*	R5P + 2 ATP ==> PRPP
59	*met*	ATP + NADPH <==> MTHF
60	*alaB*	Pyr + Glu ==> alKG + Ala
61	*avt*	2 Pyr + NADPH + Glu ==> alKG + CO2 + Val
62	*ilv*	2 Pyr + AcCoA + NADPH + Glu ==> alKG + NADH + 2 CO2 + Leu
63	*asn*	2 ATP + N + Asp ==> Asn
64	*asp*	OxA + Glu ==> alKG + Asp
65	*Lys*	di_am_pim ==> CO2 + Lys
66	*met*	SuccCoA + ATP + 2 NADPH + MTHF + Cys + Asp==> Pyr + Succ + N + Met

Table A.1 (continued)

67	*thr*	2 ATP + 2 NADPH + Asp ==> Thr
68	*ilv*	Pyr + NADPH + Glu + Thr ==> alKG + CO2 + N + Ile
69	*his*	ATP + PRPP + Gln ==> alKG + 2 NADH + His
70	*gab*	alKG + NADPH + N ==> Glu
71	*gln*	ATP + N + Glu ==> Gln
72	*pro*	ATP + 2 NADPH + Glu ==> Pro
73	*arg*	AcCoA + 4 ATP + NADPH + CO2 + N + Asp + 2 Glu ==> alKG + Fum + Ac + Arg
74	*trp*	Chor + PRPP + Gln + Ser ==> G3P + Pyr + CO2 + Glu + Trp
75	*tyr*	Chor + Glu ==> alKG + NADH + CO2 + Tyr
76	*phe*, *tyr*	Chor + Glu ==> alKG + CO2 + Phe
77	*ser*	3PG + Glu ==> alKG + NADH + Ser
78	*gly*	Ser ==> MTHF + Gly
79	*cys*	AcCoA + S + Ser ==> Ac + Cys
80	*r*ATP_Synth	5 ATP + CO2+ PRPP + 2 MTHF + 2 Asp + Gly + 2 Gln ==> 2 Fum + NADPH + 2 Glu + rATP
81	rGTP_Synth	6 ATP + CO2 + PRPP + 2 MTHF + 2 Asp + Gly + 3 Gln ==>2 Fum + NADH + NADPH + 3 Glu + rGTP
82	rCTP_Synth	ATP + Gln + rUTP ==> Glu + rCTP
83	rUTP_Synth	4 ATP + N + PRPP + Asp ==> NADH + rUTP
84	dATP_Synth	NADPH + rATP ==> dATP
85	dGTP_Synth	NADPH + rGTP ==> dGTP
86	dCTP_Synth	NADPH + rCTP ==> dCTP
87	dTTP_Synth	2 NADPH + MTHF + rUTP ==> dTTP
88	avg_FS_Synth	8.24 AcCoA + 7.24 ATP + 13.91 NADPH ==> avg_FS
89	UDPGlc_Synth	G6P + ATP ==> UDPGlc
90	CDPEth_Synth	3PG + 3 ATP + NADPH + N ==> NADH + CDPEth
91	OH_myr_ac_Synth	7 AcCoA + 6 ATP + 11 NADPH ==> OH_myr_ac
92	C14_0_FS_Synth	7 AcCoA + 6 ATP + 12 NADPH ==> C14_0_FS
93	CMP_KDO_Synth	PEP + R5P + 2 ATP ==> CMP_KDO
94	NDPHep_Synth	1.5 G6P + ATP ==> 4 NADPH + NDPHep
95	TDPGlcs_Synth	F6P + 2 ATP + N ==> TDPGlcs
96	UDP_NAG_Synth	F6P + AcCoA + ATP + Gln ==> Glu + UDP_NAG
97	UDP_NAM_Synth	PEP + NADPH + UDP_NAG ==> UDP_NAM
98	di_am_pim_Synth	Pyr + SuccCoA + ATP + 2 NADPH + Asp + Glu ==> alKG + Succ + di_am_pim
99	ADPGlc_Synth	G6P + ATP ==> ADPGlc
100	Glucose uptake	==> GLC
101	Glycerol exchange	Glyc ==>
102	*glp*	Glyc3P ==> ATP + Glyc
103	*pfkA*	F6P + ATP ==> F16P
104	*mae*	Mal ==> Pyr + NADPH + CO2
105	*Oxygen uptake*	==> O2
106	*cyc*	QuiH2 + 0.5 O2 ==> 2 H_ex
107	*Asnb*	2 ATP + Asp + Gln ==> Glu + Asn
108	*gltBD*	alKG + NADPH + Gln ==> 2 Glu
109	*Cys*	S + ASER ==> Ac + Cys
110	*ilvB*	2 Pyr ==> CO2 + ACLAC
111	*ilvC*	NADPH + ACLAC ==> DHVAL

Table A.1 (continued)

112	*ilvD*	DHVAL ==> OIVAL
113	*proB*	ATP + Glu ==> GLUP
114	*proA*	NADPH + GLUP ==> GLUGSAL
115	*aroF*	PEP + E4P ==> 3DDAH7P
116	*aroB*	3DDAH7P ==> DQT
117	*aroD*	DQT <==> DHSK
118	*aroE*	NADPH + DHSK <==> SME
119	*aroL*	ATP + SME ==> SME5P
120	*aroC*	PEP + SME5P ==> 3PSME
121	*thrA*	ATP + Asp <==> BASP
122	*asd*	2 NADPH + BASP <==> HSER
123	*metL*	ATP + HSER ==> PHSER
124	*sera*	3PG ==> NADH + PHP
125	*serC*	Glu + PHP ==> alKG + 3PSER
126	*pheA*	Chor ==> PHEN
127	*pheA2*	PHEN ==> CO2 + PHPYR
128	*trpDE*	Chor + Gln ==> Pyr + Glu + AN
129	*trpD*	PRPP + AN ==> NPRAN
130	*trpC*	NPRAN ==> CPAD5P
131	*trpC2*	CPAD5P ==> CO2 + IGP
132	*tyrA*	PHEN ==> NADH + CO2 + HPHPYR
133	*argA*	AcCoA + Glu ==> NAGLU
134	*argB*	ATP + NAGLU ==> NAGLUYP
135	*argC*	NADPH + NAGLUYP <==> NAGLUSAL
136	*argD*	Glu + NAGLUSAL <==> alKG + NAARON
137	*argE*	NAARON ==> Ac + ORN
138	*carAB*	2 ATP + CO2 + Gln ==> Glu + CAP
139	*argFI*	ORN + CAP <==> CITR
140	*argG*	2 ATP + Asp + CITR ==> ARGSUCC
141	*ilvA*	Thr ==> N + OBUT
142	*ilvBN*	Pyr + OBUT ==> CO2 + ABUT
143	*ilvC2*	NADPH + ABUT ==> DHMVA
144	*ilvD2*	DHMVA ==> OMVAL
145	*hisG*	ATP + PRPP ==> PRBATP
146	*hisI*	PRBATP ==> PRBAMP
147	*hisE*	PRBAMP ==> PRFP
148	*hisA*	PRFP ==> PRLP
149	*hisF*	Gln + PRLP ==> Glu + DIMGP
150	*hisB*	DIMGP ==> IMACP
151	*hisC*	Glu + IMACP ==> alKG + HISOLP
152	*hisB2*	HISOLP ==> HISOL
153	*metA*	SuccCoA + HSER ==> OSLHSER
154	*metB*	Cys + OSLHSER ==> Succ + LLCT
155	*metC*	LLCT ==> Pyr + N + HCYS
156	*metF*	NADH + METTHF ==> MTHF
157	*leuA*	AcCoA + OIVAL ==> CBHCAP
158	*leuCD*	CBHCAP <==> IPPMAL
159	*leuB*	IPPMAL ==> NADH + CO2 + OICAP

Table A.2 Metabolites list for *E. coli.*

No	Abbreviated name	Full Name
1	2PG	2-Phosphoglycerate
2	3DDAH7P	3-Deoxy-d-arabino heptulosonate-7-phosphate
3	3PG	3-Phosphoglycerate
4	3PSER	3-Phosphoserine
5	3PSME	3-Phosphate-shikimate
6	ABUT	2-Aceto-2-hydroxy butyrate
7	Ac	Acetate
8	AcCoA	Acetyl-CoA
9	ACLAC	Acetolactate
10	AcP	Acetyl phosphate
11	Adh	Acetaldehyde
12	ADPGlc	ADPglucose
13	Ala	Alanine
14	alKG	alpha-Ketoglutarate
15	AN	Antranilate
16	Arg	Arginine
17	ARGSUCC	L-Arginio succinate
18	ASER	O-Acetylserine
19	Asn	Asparagine
20	Asp	Aspartate
21	ATP	Adenosintriphosphate
22	avg_FS	average fatty acid
23	BASP	b-Aspartyl phosphate
24	Biomass	Biomass
25	C14_0_FS	C_14:0_Fatty_acid
26	CAP	Carbamoyl phosphate
27	CBHCAP	3-Carboxy-3-hydroxy-isocaproate
28	CDPEth	CDP ethanolamine
29	Chor	Chorismate
30	Cit	Citrate
31	CITR	L-Citrulline
32	CMP_KDO	CMP-3-deoxy-D-manno-octulosonate
33	CO2	Carbon dioxide
34	CPAD5P	1-O-Carboxyphenylamino 1-deoxyribulose-5-phosphate
35	Cys	Cysteine
36	dATP	ATP for DNA synthesis
37	dCTP	CTP for DNA synthesis
38	dGTP	GTP for DNA synthesis
39	DHAP	Dihydroxyacetone phosphate
40	DHMVA	2,3-Dihydroxy-3-methyl-valerate
41	DHSK	Dehydroshikimate
42	DHVAL	Dihydroxy-isovalerate
43	di_am_pim	Diaminopimelate
44	DIMGP	D-Erythro imidazoleglycerol-phosphate
45	DPG	Diphosphoglycerate

Table A.2 (continued)

46	DQT	3-Dehydroquinate
47	dTTP	TTP for DNA synthesis
48	E4P	D-Erythrose 4-phosphate
49	Eth	Ethanol
50	F16P	Fructose 1,6-bisphosphate
51	F6P	Fructose 6-phosphate
52	Form	Formate
53	Fum	Fumarate
54	G3P	Glyceraldehyde 3-phosphate
55	G6P	Glucose 6-phosphate
56	GLC	Glucose
57	Gln	Glutamine
58	Glu	Glutamate
59	GLUGSAL	L-Glutamate g-semialdehyde
60	GLUP	Glutamyl phosphate
61	Gly	Glycine
62	Glyc	Glycerol
63	Glyc3P	Glycerol 3-phosphate
64	Glyox	Glyoxylate
65	H_ex	External Hydrogen
66	HCYS	Homocysteine
67	His	Histidine
68	HISOL	Histidinol
69	HISOLP	L-Histidinol-phosphate
70	HPHPYR	para-Hydroxy phenyl pyruvate
71	HSER	Homoserine
72	ICit	Isocitrate
73	IGP	Indole glycerol phosphate
74	Ile	Isoleucine
75	IMACP	Imidazole acetyl-phosphate
76	IPPMAL	3-Isopropylmalate
77	KetoPGluc	2-keto-3-deoxy-D-gluconate 6-phosphate
78	Lac	Lactate
79	Leu	Leucine
80	LLCT	L-Cystathionine
81	Lys	Lysine
82	Mal	Malate
83	Met	Methionine
84	METTHF	5,10-Methylene tetrahydrofolate
85	MTHF	Methylen-Tetrahydrofolate
86	N	Nitrogen(NH4)
87	NAARON	N-a-Acetyl ornithine
88	NADH	Nicotinamide adenine dinucleotide - reduced
89	NADPH	Nicotinamide adenine dinucleotide phosphate - reduced
90	NAGLU	N-Acetyl glutamate
91	NAGLUSAL	N-Acetyl glutamate semialdehyde
92	NAGLUYP	N-Acetyl glutamyl -phosphate
93	NDPHep	NDP Heptose

Table A.2 (continued)

94	NPRAN	N-5-phosphoribosyl-antranilate
95	O2	Oxygen
96	OBUT	Oxobutyrate or 2-ketobutyrate
97	OH_myr_ac	OH myristic Acid
98	OICAP	2-Oxoisocaproate
99	OIVAL	Oxoisovalerate
100	OMVAL	Oxomethylvalerate
101	ORN	Ornithine
102	OSLHSER	O-Succinyl-l-homoserine
103	OxA	Oxaloacetate
104	PEP	Phosphoenolpyruvate
105	PGlac	6-Phospho-Gluconolactone
106	PGluc	6-Phospho-Gluconate
107	Phe	Phenylalanine
108	PHEN	Prephenate
109	PHP	3-Phosphohydroxypyruvate
110	PHPYR	Phenyl pyruvate
111	PHSER	O-Phospho-l-homoserine
112	PRBAMP	Phosphoribosyl -AMP
113	PRBATP	Phosphoribosyl-ATP
114	PRFP	Phosphoribosyl-formimino-AICAR-phosphate
115	PRLP	Phosphoribulosyl- formimino-AICAR-phosphate
116	Pro	Proline
117	PRPP	5-Phospho-alpha-D-ribose 1-diphosphate
118	Pyr	Pyruvate
119	QuiH2	Ubichinon_red
120	R5P	Ribose 5-phosphate
121	rATP	ATP for RNA synthesis
122	rCTP	CTP for RNA synthesis
123	rGTP	GTP for RNA synthesis
124	Rl5P	Ribulose 5-phosphate
125	rUTP	UTP for RNA synthesis
126	S	Sulfur(SO4)
127	S7P	Sedoheptulose 7-phosphate
128	Ser	Serine
129	SME	Shikimate
130	SME5P	Shikimate-5-phosphate
131	Succ	Succinate
132	SuccCoA	Succinyl-CoA
133	TDPGlcs	TDP-glucosamine
134	Thr	Threonine
135	Trp	Tryptophan
136	Tyr	Tyrosine
137	UDP_NAG	UDP acetylglucosamine
138	UDP_NAM	UDP N-acetylmuramic acid
139	UDPGlc	UDP glucose
140	Val	Valine
141	X5P	Xylolose-5-Phosphate

Appendix B

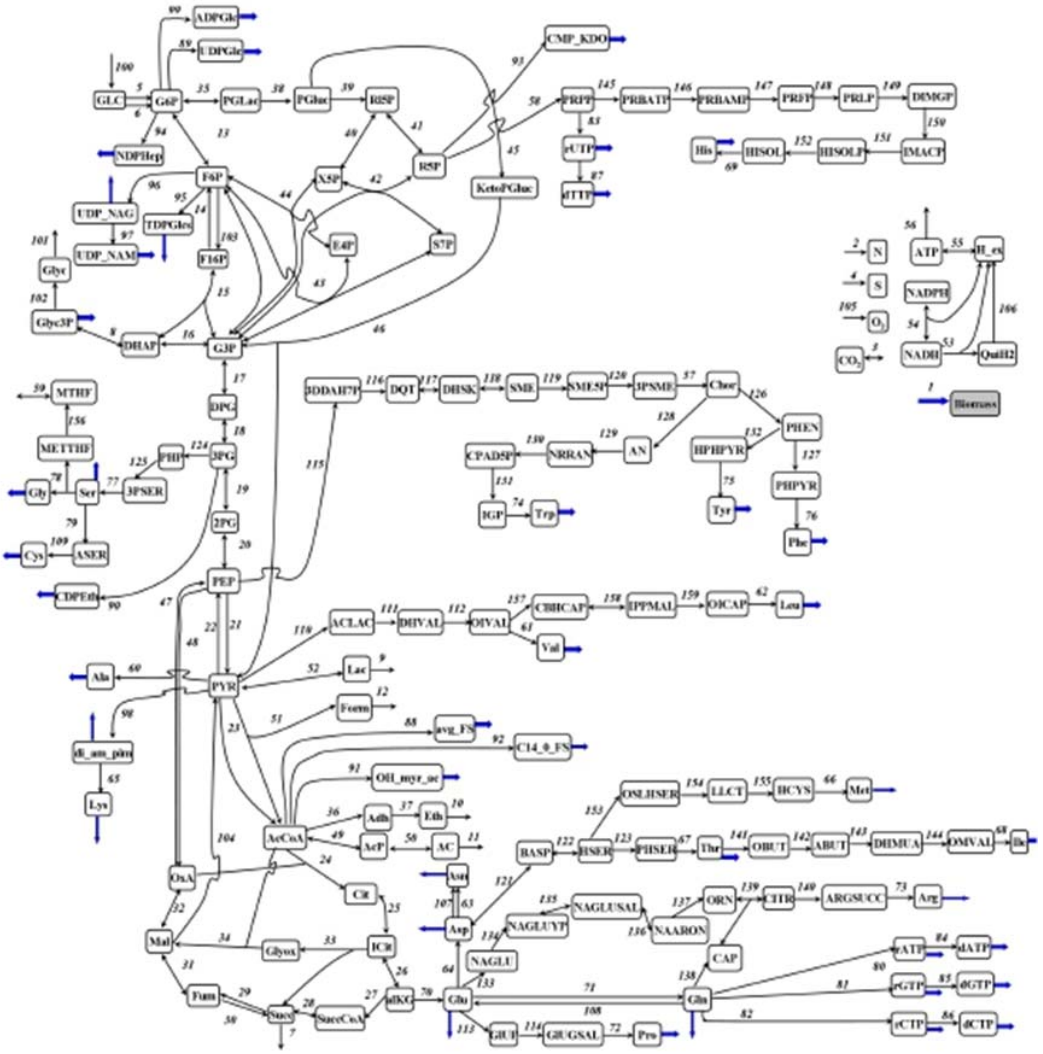


Figure B.1: Metabolic network map for *E. coli*. Details of the metabolic reactions and metabolites of *E. coli* are shown in Table A.1.

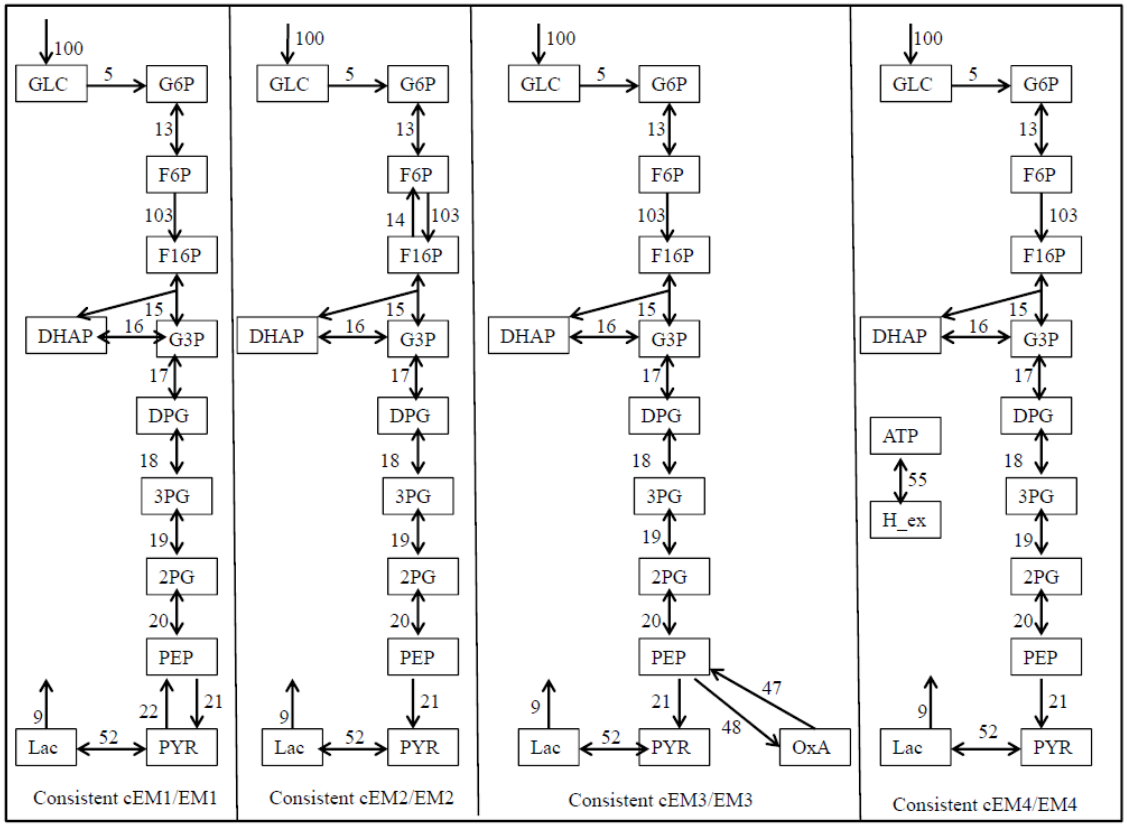


Figure B.2: Metabolic pathways for the 4 consistent EMs or cEMs for Model-I. Details of the metabolic reactions and metabolites are shown in Table A.1. (Numerical values indicate the reaction number, details in Figure B.1).

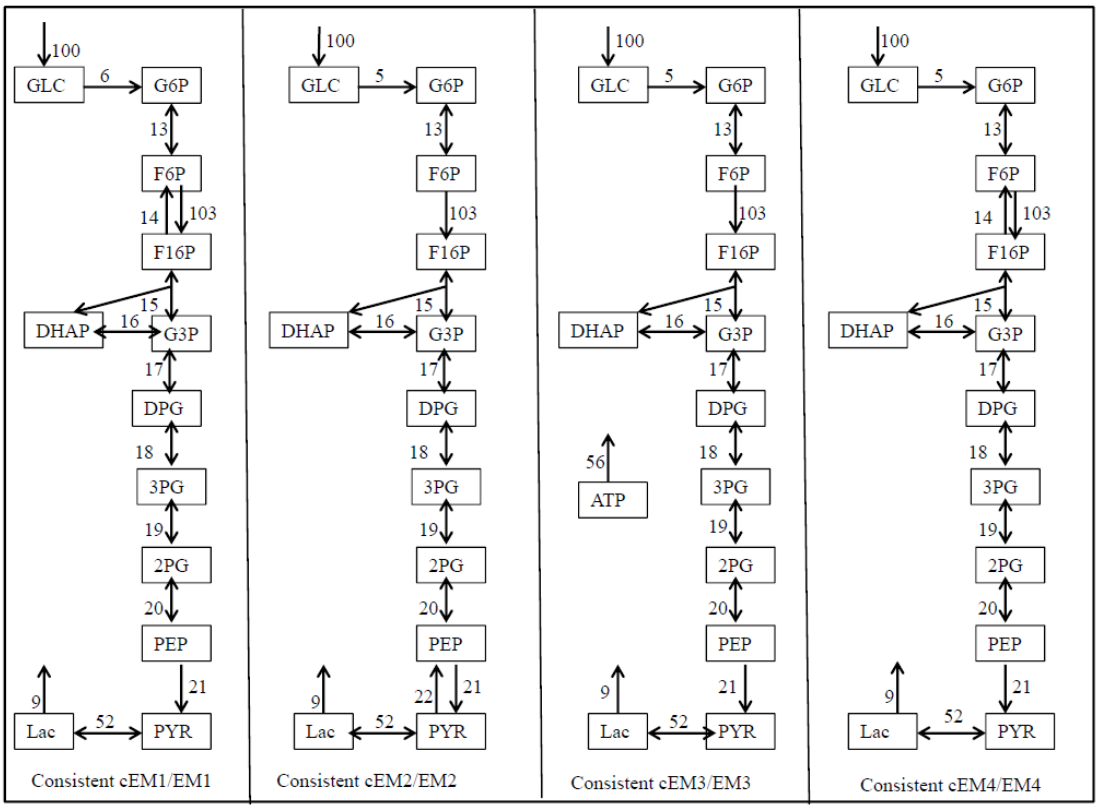


Figure B.3: Metabolic pathways for the 12 consistent EMs or cEMs for Model-II. Details of the metabolic reactions and metabolites are shown in Table A.1.

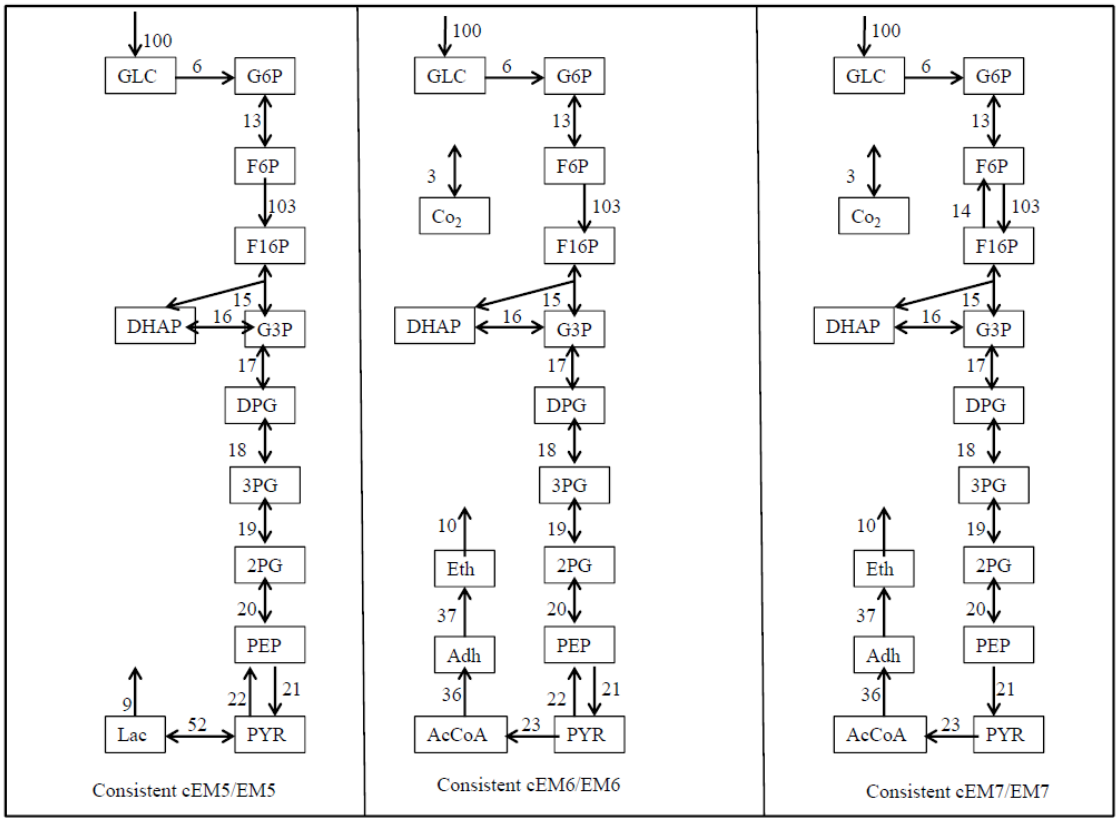


Figure B.3 (Continued)

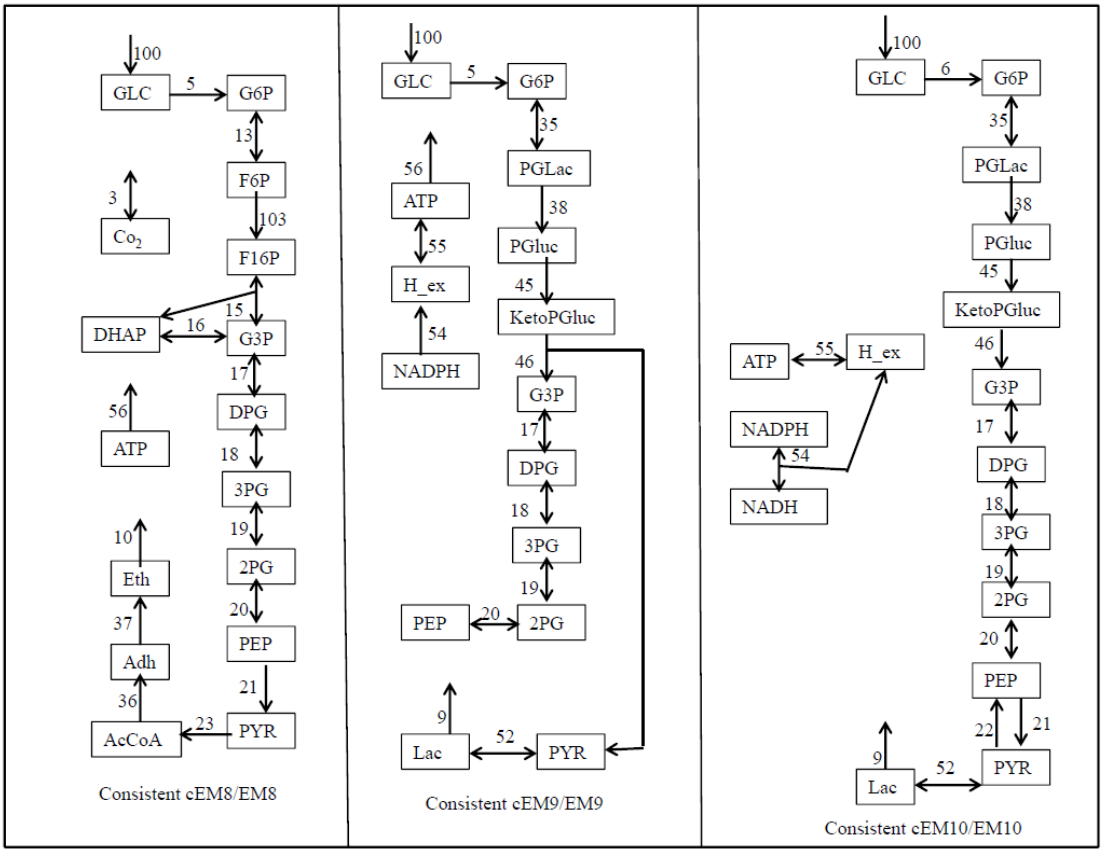


Figure B.3 (Continued)

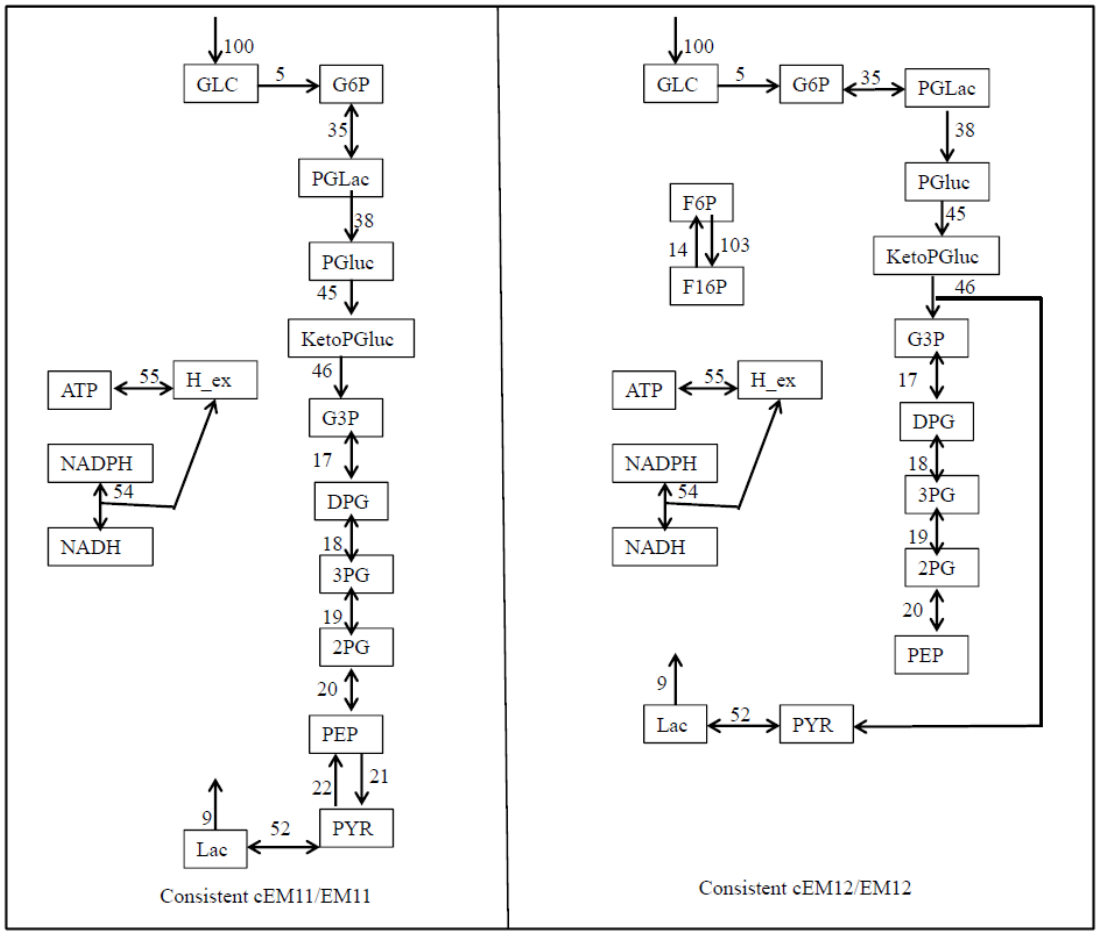


Figure B.3 (Continued)